THE UNIVERSITY
of EDINBURGH

Application of Multivariate Statistics and Machine Learning to Phenotypic Imaging and Chemical High-Content Data

Author Jan Wildenhain

THE UNIVERSITY
of EDINBURGH

Thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy
to the
University of Edinburgh — 2016

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, either in whole or in part, in any previous application for a degree. Except where otherwise acknowledged, the work presented is entirely my own.

Jan Wildenhain

September 2016

Abstract

Image-based high-content screens (HCS) hold tremendous promise for cell-based phenotypic screens. Challenges related to HCS include not only storage and management of data, but critical analysis of the complex image-based data. I implemented a data storage and screen management framework and developed approaches for data analysis of a number high-content microscopy screen formats. I visualized and analysed pilot screens to develop a robust multi-parametric assay for the identification of genes involved in DNA damage repair in HeLa cells. Further, I developed and implemented new approaches for image processing and screen data normalization. My analyses revealed that the ubiquitin ligase RNF8 plays a central role in DNA-damage response and that a related ubiquitin ligase RNF168 causes the cellular and developmental phenotypes characteristic for the RIDDLE syndrome. My approaches also uncovered a role for the MMS22L-TONSL complex in DSB repair and its role in the recombination-dependent repair of stalled or collapsed replication forks.

The discovery of novel bioactive molecules is a challenge because the fraction of active candidate molecules is usually small and confounded by noise in experimental readouts. Cheminformatics can improve robustness of chemical high-throughput screens and functional genomics data sets by taking structure-activity relationships into account. I applied statistics, machine learning and cheminformatics

to different data sets to discern novel bioactive compounds. I showed that phenothiazines and apomorphines are regulators for cell differentiation in murine embryonic stem cells. Further, I pioneered computational methods for the identification of structural features that influence the degradation and retention of compounds in the nematode *C. elegans*. I used chemoinformatics to assemble a comprehensive screening library of previously approved drugs for redeployment in new bioassays. A combination of chemical genetic interactions, cheminformatics and machine learning allowed me to predict novel synergistic antifungal small molecule combinations from sensitized screens with the drug library. In another study on the biological effects of commonly prescribed psychoactive compounds, I discovered a strong link between lipophilicity and bioactivity of compounds in yeast and unexpected off-target effects that could account for unwanted side effects in humans. I also investigated structure-activity relationships and assessed the chemical diversity of a compound collection that was used to probe chemical-genetic interactions in yeast. Finally, I have made these methods and tools available to the scientific community, including an open source software package called MolClass that allows researchers to make predictions about bioactivity of small molecules based on their chemical structure.

Acknowledgements

First of all, I would like to thank Mike Tyers for his advice, support and sponsorship over the last 10 years. We had lively scientific, philosophical and political discussions over a wee dram of whiskey. I am grateful that he gave me the freedom and opportunity to work and collaborate with a wide variety of scientists. I would also like to thank some of my collaborators over the years for their great teamwork, support, openness and dedication: Phedias, Andrew, Maureen, Nadine, Stephanie, Sonam, Steven, Giora, Tracy, Andrew and Jarkko. I would like to thank David B., John R., Mark E., Ricardo A. and Jeff S. whom in their ways provided me with mentorship, wise words and good memories to cherish. I would like to thank Sander Granneman to support my PhD by research application, my examiners Chris Bakal, Andrew Hopkins and Ian Overton for their critical suggestions. Also not to forget Simon Langdon who was acting as non-examining chair providing us with a ride to Dundee. Finally, my friends who always encourage and supported me in times where I had doubts about making the right choices like Angel & Lutz, Lauri, Aditi & Richard, Juri & Susi, Stefan & Melli, Nhan, Sasha, Mark and Alex. Special thanks go to Hina for hosting me through the last year in Edinburgh. Most importantly and not to forget, I would like to thank my parents and my co-scientist and partner Michaela for their unconditional trust and support. I'll always cherish my time and colleagues at the University of Edinburgh, which is a marvellous place to work and who knows maybe our ways will cross again.

Contents

List of Tables

List of Figures

List of Acronyms

5-hmC 5-Hydroxymethylcytosine

5-mC 5-Methylcytosine

ABC ATP Binding Cassette

ADMET Absorption, Distribution, Metabolism, Excretion and Toxicity

AI Artificial Intelligence

AIC Akaike Information Criterion

AID PubChem BioAssay IDentifier

AF-555 Alexa Fluor 555

ANOVA ANalysis Of VAriance

API's Application Programming Interfaces

AR AutoRegressive

ATM Ataxia telangiectasia mutated

AUC Area Under the Curve

BAG3 BCL2 associated athanogene 3

BRCA1 BReast CAncer 1 tumor suppressor gene

CAS Chemical Abstracts Service

CellH5 HDF5 data format for cell-based assays

CCNA2 Cyclin A2

CDF Cumulative Density Function

CHX cycloheximide

CNS Central Nervous System

CORDY cordycepin

D3 Data-Driven Documents

DAD Diode Array Detector

DAPI 4',6-diamidino-2-phenylindole

DAVID Database for Annotation, Visualization and Integrated Discovery

DCMU 3-(3,4-dichlorophenyl)-1,1-dimethylurea

DMSO DiMethyl SulfOxide

DNA DeoxyriboNucleic Acid

DDR DNA Damage Repair

DSB Double Strand Break

ECFP4 Enhanced Connectivity FingerPrints of length 4

ECFP Enhanced Connectivity FingerPrints

ECM ExtraCellular Matrix

EGFP Enhanced Green Fluorescent Protein

EMBL European Molecular Biology Laboratory

ER Entity-Relationship

esiRNA endoribonuclease-prepared siRNA

FACS Fluorescence-Activated Cell Sorting

FCFP Functional-Class Fingerprints

FDA U.S. Food and Drug Administration

FP False Positive

FN False Negative

FFT Fast Fourier Transform

G_1 Growth phase 1

G_2 Growth phase 2

GFP Green Fluorescent Protein

GPCR G-Protein-Coupled Receptor

GO Gene Ontology

GWAS Genome-Wide Association Study

HA Hydrogen bond Acceptors

HD Hydrogen bond Donors

HCA High Content Analysis

HCS High Content Screen

HCF Host Cell Factor C1

HCI High-content imaging

HeLa Henrietta Lacks cervical cancer

HEK Human Embryonic Kidney

hESC human Embryonic Stem Cells

HIP HaploInsufficiency Profiling

HMM Hidden Markov Models

HOP HOmozygous deletion Profiling

HPLC High Performance Liquid Chromatography

HPRD Human Protein Reference Database

HR Homologous Recombination

HTS High Throughput Screen

HTML5 Hyper Text Markup Language 5

HVEM Tumor necrosis factor receptor superfamily member 14

IAA Indole-3-Acetic Acid

IHOP Information Hyperlinked Over Proteins

ICP4 *herpes virus* transcriptional regulator

IQR Interquartile range

KR Klekota-Roth

KS Kolmogoroff-Smirnov

KNN K-Nearest Neighbours

LIG4 Ligase IV

LC-MS Liquid Chromatography-Mass Spectrometry

LMT Logistic Model Tree

LOPAC Library of Pharmacologically Active Compounds

LOWESS LOcally WEighted Scatterplot Smoothing

MACCS Molecular ACCess System

MAD Median Absolute Deviation

MD Mahalanobis Distance

MDDR Molecular Drug Data Report

MDR Multi-Drug Resistant

MDS MultiDimensional Scaling

ML Machine learning

MMP Matrix MetalloProteinases

MMS22 Methyl Methanesulfonate Sensitivity 22

MOA Mode Of Action

MTT 3-[4,5-dimethylthiazol-2-yl]-2,5-diphenyltetrazolium bromide

MW Molecular Weight

MYC Myelocytomatosis viral oncogene

NANOG Nanog Homeobox transcription regulator

NB Naive Bayes

NC Nitrogen Count

NCBI National Center for Biotechnology Information

NCTRL Negative ConTRoL

NGS Next-Generation Sequencing

NIH National Institute of Health

NLS Nonlinear Least Squares

NMR Nuclear Magnetic Resonance

NR Number of Rings

NSA Nucleic Size Area

NSCs Neural Stem Cells

OC Oxygen Count

OCT4 Octamer-binding transcription factor 4

OME Open Microscopy Environment

OMIM Online Mendelian Inheritance in Man

ORF Open Reading Frame

ORF28 *Herpes virus* DNA polymerase catalytic subunit

QSAR Quantitative Structure Activity Relationship

PAD Previously Approved Drugs

PDB Protein Data Bank

PCA Principal Component Analysis

PCNA Proliferating Cell Nuclear Antigen

PDR Pleiotropic Drug Resistance

PHP Hypertext Preprocessor

PI Propidium Iodide

PLK1 Polo Like Kinase 1

PRKAB1 PRotein Kinase AMP-activated non-catalytic subunit Beta 1

PSA Polar Surface Area

PSI-MI Proteomics Standards Initiative - Molecular Interaction

PPV Positive Predictive Value

QC Quality Control

RAD51 RADiation repair gene 51

RB Rotatable Bonds

REST REpresentational State Transfer

RF Random Forest

RIDDLE Radiosensitivity, Immunodeficiency, Dysmorphic, Difficulties LEarning

RNA RiboNucleic Acid

RNAi RNA interference

RND Resistance-Nodulation-Division

RNF168 E3 ubiquitin-protein ligase, Ring Finger Protein 168

RNF8 E3 ubiquitin-protein ligase, Ring Finger Protein 8

ROC Receiver Operating Characteristic

RRM2 Ribonucleotide Reductase M2

RSCF siRNA which is not processed by the RISC machinery

SC Sulphur Count

SAR Structure Activity Relationship

SD Standard Deviation

SDS Sodium Dodecyl Sulfate

siRNA small interfering RNA

SMB Server Message Block

SMILES Simplified Molecular-Input Line-Entry system

SMOTE Synthetic Minority Over-sampling TEchnique

SP1 Sp1 transcription factor

ssDNA single-stranded DNA

SOM Self-Organizing Maps

SONAR Second Order Neighbour Activity Response

SSMD Strictly Standardised Mean Difference

STRING Search Tool for Recurring Instances of Neighbouring Genes

SVGs Scalable Vector Graphics

SVM Support Vector Machine

TIN2 (TRF1)-Interacting Nuclear factor 2

TN True Negatives

TP True Positives

TP53BP1 Tumor Protein p53 Binding Protein 1

Trp Tryptophan

TSA Total Spot Area

TSN Total Spot Number

TSI Total spot intensity

U2OS human osteosarcoma cell line

UV UltraViolet

VACV *VACcinia Virus*

VACVPol1 *Vaccinia virus* polymerase siRNA pool 1

VACVPol2 *Vaccinia virus* polymerase siRNA pool 2

VP16 Virion protein involved in morphogenesis

WEKA Waikato Environment for Knowledge Analysis

WHO World Health Organisation

WOMBAT WOrld of Molecular BioAcTivity

XML eXtensible Markup Language

YBCL Yeast Bioactive Compound Library

Chapter 1

Background

This chapter will give an overview of the key developments in High Content Screen (HCS) and chemical High Throughput Screen (HTS) . The work submitted here started in 2005 and I will therefore focus on academic papers published until 2005 that guided my choices of methods to apply to high content and high throughput screening data. HTS and high-content imaging High-content imaging (HCI) will be introduced and I will outline the history of computational approaches for the analysis of such data. Specifically, I will focus on the use of machine learning in image analysis and the history and translational value of cheminformatics.

1.1 High Content Imaging Screens

High content screening is a multi-parametric acquisition technique that uses automated digital microscopy to monitor activation, inhibition or perturbation of a protein or biological process by a small molecule or other bioactive agent. The first HCS platform was developed in the early 1990s at the University of

Pittsburgh (Farkas et al. 1993). HCS was primarily based on the light microscope, an instrument developed in the 17th century. In the early days, scientists relied on microscopy to observe qualitative features in biological samples. HCS represents the modern form of discovery-driven research which, in contrast to the early days, uses computerised systems for feature detection and aims to quantify those patterns simultaneously in hundreds to thousands of samples. With the rise of functional genomics and the emergence of affordable robotics and compound libraries for academic institutions, interest in HCS has increased dramatically over the past two decades. The early 1990s also brought a breakthrough in new fluorescence-based reagents that could be expressed in different model systems as recombinant proteins. The first fluorescent protein gene construct came from *Aequoria victoria* (Chalfie et al. 1994) and two years later the first engineered optimised fluorescent proteins reached the scientific community (Heim and Tsien 1996). Various fluorescent proteins with different emission spectra are available now and enable researchers to monitor various parameters in cells simultaneously. The parallel development of affordable high resolution digital cameras enabled instrumentation for multicoloured and multimodal microscopy to be developed by several companies, including Evotec (now part of PerkinElmer), Cellomics (now part of Thermo Fisher Scientific) and Molecular Devices. The Human Genome Sequencing Project (Lander et al. 2001), set the stage for large-scale experimental designs to interrogate the function of the 20,000 to 25,000 human genes. The discovery of RNA interference (RNAi) (Fire et al. 1998) opened a new avenue for loss of function studies in different model organisms and human cell lines. These new approaches to understanding biological functions and pathways in a systematic fashion allowed HCS to be applied to in-depth studies of multiple proteins and/or functions simultaneously. Different image-based readouts can thus measure the impact of genetic or chemical perturbation on different cellular processes like cell cycle, motility, apoptosis and DNA damage repair. HCS has seen a major increase in data acquisition rates and the vast

amount of data requires new tools to archive, mine and display the complex imaging data. Crucially, this data deluge has created a need for rigorous statistical methods to detect and correct data acquisition errors, assay variability and data inconsistencies to ensure robust readouts for hit confirmation and follow up studies. Most software solutions for HCS measurements use directed algorithms to find patterns in cell-based screens such as nuclei, intracellular foci, and cell or other organelle boundaries. For most assays, optimal screening parameters are unknown at the start and need to be tuned for discovery. Unbiased pattern recognition and machine learning approaches are powerful tools to explore the contribution of different variables and parameter settings.

1.1.1 Computational Approaches for HCS Data

Procedures for statistical analysis of HCS data have mostly been adopted from High Throughput Screens (HTS), developed in the pharmaceutical industry (Malo et al. 2006; Dragiev, Nadon, and Makarenkov 2011). These analysis methods include general screen quality metrics such as the Z-Factor (J.-H. Zhang, Chung, and Oldenburg 1999), Strictly Standardised Mean Difference (SSMD), median normalisation, B-Score and Z-Score to reliably identify hits (Brideau et al. 2003). In HCSs, these metrics are commonly applied to the mean of a single parameter of interest determined for every single well in a study. The data obtained from classical single parameter HTS are typically measurements based on enzyme activity or other protein-associated parameters in the case of biochemical screens, or cell viability for cell-based screens. These numbers represent an average of the cell populations in each well. In HCS, images are acquired that contain several signals, each marking a different cellular component. Different features can be be extracted for these objects, including size, intensity and texture. In addition, HCS allows for extraction of single cell measurements to monitor populations

of cells. The size of the cell population becomes very important to reliably determine parameters. In a population, cells are likely to be asynchronous because cell processes are regulated on multiple levels: transcription (e.g. methylation, histone modification, transcription factor binding), translation, post-translational modifications, ubiquitination, pathway signalling cascades and most importantly, the cell cycle. Many of these processes are of stochastic nature (Perkins and Swain 2009). Within a well, a cell population might show the same pattern for one marker and a different localisation pattern for a second marker. This results in a multi-dimensional trajectory state for cell populations that is only limited by the diversity of readouts that can be observed given the assay design. Working with multiple parameters also raises the issue of variable interdependency. It is very common for parameters to be dependent. The assessment of relationships between variables can be performed with scatterplots that visualise the distribution of parameters and the correlations between them. Correlation between parameters can represent dependencies, independence or redundancies and their relationship needs to be explored in detail (Muller 1981).

To detect more subtle dependencies, multivariate techniques that include Principal Component Analysis (PCA) (Bro and Smilde 2014; Hotelling 1933), multidimensional scaling (Shepard 1980; Torgerson 1952) and canonical correlations (Bartlett 1941; Kim, J. Kittler, and Cipolla 2007) can be applied. High content data generally consists of a number of variables (feature vectors) recorded on a number of individual cells, fields or objects. There will be n objects and q features, and the data can be arranged in an $n \times q$ data matrix. Biplots visualise parameter relationships, after calculating the principal components from a data matrix (Gabriel 1971). PCA is a powerful exploratory technique that requires some mathematical knowledge (Bro and Smilde 2014). Principal components are the orthogonal combinations that maximise the observed variance in the data. For q variables, there are q principal components: the first component contains

the maximum variance of a linear combination of parameters within the dataset; the second component contains the maximal variance among linear combinations orthogonal to the first and so on. The input variables can be displayed as loadings, represented as a leading arrow where length and direction are proportional to the weight of the corresponding variable in the principal component. PCA often serves as a first processing step to feed into additional analysis methods, such as graphical Gaussian models (D. K. Singh et al. 2010; Slacka et al. 2008), multidimensional scaling (D. K. Singh et al. 2010; Slacka et al. 2008), clustering (Bakal et al. 2007; Qiu et al. 2011) or machine learning (Bakal et al. 2007; D. K. Singh et al. 2010), to segregate relevant information from background noise. While some published HCS studies apply these sophisticated methodologies, the majority (80%) restricts data analysis to single property evaluation (S. Singh, A E Carpenter, and Genovesio 2014).

In summary, the on-going development of novel technologies for HCS and evermore sophisticated assay formats have positioned statistical analysis of HCS as a critical and actively evolving field. The development and application of statistical methods for HCS has been the main focus of my research at the University of Edinburgh during the past 10 years.

1.1.2 Machine Learning in HCS Data Analysis

If metrics that we derive from images are too complex and outcomes are closely related, it is beneficial to use methods that can be trained by exploring subsets of data and assigning outcome instead of optimising parameters by trial and error. Machine learning (ML) emulates human learning, reasoning and decision making, using techniques that combine Artificial Intelligence (AI) algorithms, data mining and statistics. In the context of image processing, approaches of computer vision and ML for the analysis of HCS data is frequently termed High Content Analysis

(HCA). ML generally proceeds in two phases. In the training phase a collection of data samples is used to build or improve the ML algorithm by learning the inherent data structure and relationships. In the second phase, the algorithm is applied to new data samples that were not observed in the training phase. The overall goal is to be able to generalise from few samples and make accurate predictions about future observations. There are two classes of ML algorithms, unsupervised and supervised. The most common unsupervised learning technique is clustering. The exploration of the data is done by finding hidden patterns in multidimensional data using a similarity metric and constants of expected outcomes like a maximum cluster size. Prominent methods are hierarchical clustering, k-means clustering, Self-Organizing Maps (SOM) and Hidden Markov Models (HMM)s. Supervised learning algorithms generally require categorised data samples that are used to train an AI algorithm which will then be able to assign unseen test data into the predefined categories. In HCS, given a new input image, such an algorithm could be used to predict the different phenotypes with a specific degree of certainty.

The quality of a learner is measured by how well it balances the bias-variance tradeoff. Consider a model with two possible outcomes, cells are either mitotic or non-mitotic. The feature vector Q is used by the classification algorithm to optimise parameters θ to recapitulate outcome assignments made by the scientist. Imagine we have two variables that describe the intensity q_i and the area q_{i+1} of a cell nucleus as elements of Q. A reasonable hypothesis would be to assume that a cell with low nucleic intensity and a large area is non-mitotic and cells with a small area and high intensity are mitotic. To build a model that can predict the cell cycle state of the cell with high accuracy and precision we can propose the following:

$$Y = f(Q, \theta) + \epsilon$$

with a function $f(Q, \theta)$ that maps the inputs Q to the output Y and the deterministic relationship between Q and Y. The parameters of θ are hidden variables tuned (learned) by the function. ϵ represents a random variable that captures the independent variation not explained by Q and θ and is referred to as error with normally distributed values $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$. The estimate E from the predictive model $\hat{f}(Q, \theta)$ will have an expected squared prediction error at point q:

$$Err(q) = E[(Y - \hat{f}((q, \theta))^2]$$

This error may then be decomposed into bias and variance components:

$$Err(q) = \Big(E[\hat{f}(q, \theta)] - f(q)\Big)^2 + E\Big[\Big(\hat{f}(q, \theta) - E[\hat{f}(q, \theta)]\Big)^2\Big] + \sigma_e^2$$

The irreducible error is the noise term that cannot be reduced by any model. Given a perfect model and infinite data to calibrate it, an AI algorithm should be able to reduce both, bias and variance to zero. Given imperfect models and finite data, there is a tradeoff between minimising bias and variance.

The power of ML becomes obvious when a subset of images can be used to correctly classify new data. Image classification by supervised ML can be performed at the level of pixels, rasterized images and segmented cell objects. Pixel classifiers use local pixel neighbourhoods to learn how to separate foreground from background and whether pixels belong to certain objects (Sommer, Straehle, et al. 2011). Similarly, rasterising slices images into rectangular patches and learns the phenotypic distances between those patches for classification. Both approaches are commonly used to quantify differences in cell based screening but

are not used for single cell analysis. Object segmentation approaches usually use several fluorescent markers to stain the object of interest. Each object can be described by quantitative features that form the basis for a classifier that can distinguish them. For example, a raw pixel approach is unsuitable as it withholds information of spatial and spectral patterns in relationship to the orientation of cells and their neighbourhoods. In segmented approaches, object orientation and location can be identified and the extracted features, such as texture, can be optimised to contain such information. A simple example for a texture is an adjacent distribution of pixel intensities. The extracted features could represent mean and standard deviation. More advanced features are pixel-pixel co-occurrence patters (Haralick 1979). Many powerful morphometric features are abstract representations that are difficult to relate to a phenotype but provide powerful inputs to learning algorithms. The collection of many features improves correct classification, yet with each additional feature, the dimensional complexity increases, leading to a need for larger training and testing datasets to control bias. This is referred to as the 'curse of dimensionality' (Bellman 1961) and can be addressed with algorithms that reduce dimensionality, such as PCA, introduced in the previous section. There are generative and discrete algorithmic approaches to supervised ML. Generative methods model statistical distributions to reflect the data objects derived from the training data. These models can be deployed to generate missing data points that can be beneficial for certain approaches such as HMMs. Discriminant approaches model decision boundaries between different classes such as hyperplanes, and are employed by Support Vector Machine (SVM) algorithms. Linear discriminate models are very robust against noise but will have difficulties to separate classes if they are distributed in complex patterns. Algorithms can also be combined to build ensembles that make classification decisions based on a majority vote or a regression model. Of utmost importance with computer algorithms is the dependence on highly reproducible imaging data. ML algorithms are designed to generalise from examples, but

can only generalise from the variability that was present in the training data. Slight changes in the focal plane and dispersed intensity in images that are not noticeable to the human eye, can introduce levels of variability that lead to systematic misclassification. Further, cell densities or differences in low-level fluorescent features can compromise ML methods. The design and selection of the screening assay and features need careful *a priori* considerations. ML has long been applied to image classification problems, commonly known as computer vision or pattern recognition. Sophisticated algorithms have been used in industry for several decades. For example, the first handwriting detection algorithm was developed by Guberman in 1962 in the USSR and is now a standard procedure used in letter sorting machines around the world. In cell biology, there are only few publications available before 2005 and they focused on classification tasks to correctly identify cell organelles, such as Golgi vesicles, mitochondria, nucleoli and actin filaments (Boland and Murphy 2001; Conrad et al. 2004). An extensive overview of good practices for HCS and HCA can be found in (Buchser et al. 2014).

1.2 History of Cheminformatics

Chemistry, with its reliance on documentation was an early adopter of computer supported storage and retrieval of information. First studies in the late 1950s focused on methods for searching databases of small molecules to predict biological and chemical properties. Most of the early research and development was carried out by the Chemical Abstracts Service (CAS). One of the first approaches published was a substructure search algorithm that was based on graph theory (Ray and Kirsch 1957). Graph theory provides an almost natural transition to 2D chemical structures were nodes represent different atoms and edges represent the chemical bonds. Graph isomorphisms can be employed to test if two graphs

are equivalent, if a graph is a subgraph of another one and to find the maximum common subgraph between two graphs using 2D molecule representations. An essential requirement was the ability to provide an unambiguous and unique connection table for each molecule that was met with the development of the Morgan algorithm, which is still used today (Morgan 1965). The ongoing efforts to represent molecules electronically were accompanied by similar attempts to access and construct databases of chemical reactions. The concept of reaction indexing and computer-aided synthesis design, where an algorithm suggests sequences of reactions that could result in a desired synthetic compound, were conceptualised around the same time (Vleduts 1963). The first attempt to predict Structure Activity Relationship (SAR) was developed using Hammett substituent constants (Hammett 1935) and partitioning coefficients to correlate the biological activity of phenoxyacetic acids (Hansch et al. 1962). This work is very important because it introduces a classification model to predict activity from an input of constants and coefficients as discussed in the previous section. It also introduced the 1-octanol partition coefficient (LogP), a chemical property that is now widely used in chemical and pharmaceutical sciences. Further, it can be seen as the first Quantitative Structure Activity Relationship (QSAR) model used for biological activity discovery and for years it was referred to as the Hansch analysis. Another important QSAR approach was substructural analysis (Cramer et al., 1974) which was the first drug discovery application of ML, two decades before ML became of interest to the cheminformatics community.

The broad strategy to organise chemical information was laid out in the early 1980's, focused on efficient and fast information retrieval using complex search methods (Lynch, Barnard, and Welford 1981). The breakthrough for cheminformatics, the computational processing of small molecule information and property

calculation, began in the late 1980s with the introduction of Simplified Molecular-Input Line-Entry system (SMILES) and the arrival of the first personal computers (Weininger 1988). SMILES were advantageous to connection tables since they were represented as text and more readable. The early 1990s lead to extended research in predictive calculations of compound properties, for example solubility constants or compound charges and to pinpoint the basis of SARs. The inceased availability of protein crystal structure information supported this trend further. Since the late 1990s research has been focussed on accessibility, performance, scalability and application of the developed algorithms. Nowadays, scientists process ten thousands of chemical compounds and analyse and annotate the relationship of their structures with protein binding (B. Chen, Wild, and Guha 2009; Gaulton et al. 2012; Irwin et al. 2012), therapeutic and adverse responses (Duran-Frigola, Rossell, and Aloy 2014; M Kuhn et al. 2011) and protein or genomic interaction data (Cheng, C. Liu, et al. 2012; Westen and Overington 2013; Cheng and Zhao 2014). With the constantly increasing size of public and private data repositories, statistical normalisation and machine learning methods will continue to gain importance for data analysis.

Academia has historically contributed to exploring fundamental aspects of disease by employing and exploring new methods to understand genetic and biochemical processes that are perturbed in human diseases. High throughput screening, a key technique to assess the activity of individual compounds across large collections of small molecules, was limited by financial and infrastructure restrictions in academia. With a recent emphasis on translational science by government and funding agencies and the reduced cost of robotic liquid handling and compound libraries, academic labs have made substantial forays into drug discovery over the past decade. Approved drugs currently only target 120 protein domain families and a quarter of drugs are focused on G-Protein-Coupled Receptor (GPCR)'s (Overington, Al-Lazikani, and A. L. Hopkins 2006). The number of unique

molecules that are prescribed as drugs is close to 1,400 (Law et al. 2014). The theoretical chemical space of small molecules up to 15 atoms consisting of C, N, O, F, Cl and S is estimated to be 28 billion compounds with an average molecular weight of around 207 Da (Reymond et al. 2010). Every HTS campaign thus covers only a small fraction of chemical space against a specific biological function. Data sharing is therefore imperative to explore chemical space effectively. Freely available repositories such as ChEMBL (Gaulton et al. 2012) and PubChem (B. Chen, Wild, and Guha 2009) have only been established in the last few years. These resources link small molecule descriptors and information about biological assays.

To be able to build large and more diverse datasets for computational models, joint collaborative efforts are needed to cover large chemical space (A. L. Hopkins and Bickerton 2010) and to develop better lead descriptors (Bickerton et al. 2012). Efforts to build large-scale robust prediction models are most certainly enabled by resources such as ChEMBL and Toxbank (Kohonen et al. 2013), although these resources are inherently biased by input chemistries. Some authors argue that small molecule similarity should be defined based on similarity of their biological activity (Petrone et al. 2012). It has been demonstrated that biases can be overcome with the integration of other datasets such as protein sequence similarities (Keiser, Setola, et al. 2009; Yabuuchi et al. 2011), network information (Jia et al. 2009; M Kuhn et al. 2011; Yildirim et al. 2007), adverse effect similarities (Campillos et al. 2008; Michael Kuhn et al. 2013) and gene expression data (Bernardo et al. 2005; Lamb 2006).

The majority of studies that employ ML techniques showcase proof-of-concept examples that discover strong associations with well-studied pathways and known mode of action. The use of unsupervised and supervised machine learning algorithms will inevitably increase in the near future since the amount of data and its regulatory interconnectivity are more complex than ever before.

1.3 Thesis Contributions

This thesis contributes to the field of data driven research in HCS, chemical genetics and HTS. An ongoing challenge is to choose the appropriate techniques required to analyse experimental data acquired in large biological screening campaigns. The ultimate goal of bioinformatics is to increase our understanding of biological processes by means of developing and applying computational methods. Those methods are broadly distributed across the fields of engineering, statistics, mathematics, computer science and physics. A selected set of techniques from statistics, pattern recognition, data mining, cheminformatics and machine learning have been applied in this work. This thesis aims to highlight the complexity of large scale experimental designs and suggests computational methods and workflows that can be applied to future experiments of similar scope.

1.4 Thesis Structure

The following chapters focus on selected projects I have worked on over the past 10 years. Chapter 2 focusses on the analysis of HCS data beginning with the initial stages of data acquisition. It describes the problems we faced and the solutions that were developed. It will also highlight the results and their impact after publication. Chapter 3 will discuss my work in the field of chemical genetics and chemical HTS with data generated in several species and cell lines. All of this work was derived using different experimental platforms and faced different computational challenges related to experimental platform, data acquisition, normalisation and hit selection. Chapter 4 will introduce MolClass, an application I developed to evaluate and classify small molecules from different types of experiments. The performance of different molecule representations and ML methods will be compared and evaluated. Chapter 5 presents general

conclusions based on the published work and will discuss steps necessary to keep a competitive edge in future directions.

Chapter 2

Development of Statistical Methods for High Content Data

Image-based high-content screens (HCS) hold tremendous promise for the discovery of new biological functions and bioactive agents using cell-based phenotypic readouts. Challenges related to HCS include storage, management and comprehensive analysis of the large and complex image-based data sets. I have been actively involved in HCS assay development and developed data analysis procedures to support scientific projects performed at the University of Toronto and the University of Edinburgh. Notably, my work has contributed to several publications which will be described in more detail in this chapter (Kolas et al. 2007; Stewart, Panier, et al. 2009; O'Donnell et al. 2010; Beard et al. 2014).

The main focus of this chapter is the development of a large-scale High Content Screen (HCS) to discover novel genes involved in DeoxyriboNucleic Acid (DNA) damage repair. This work has been performed under the guidance of Daniel Durocher and I worked closely together with two members of his laboratory, Nadine Kolas and Jarkko Ylanko. My contributions to this work have been critical in

the early stages for optimising the screening assay and later on for performing the data analysis. I have been involved in making decisions regarding the choice of the fluorescent markers, positive and negative controls, image acquisition and assay design. I developed and implemented approaches for image processing and screen data normalisation, built a data storage and screen management framework and developed and implemented approaches for data analysis of different high-content microscopy screen formats. My visualisation and analysis of the pilot screens enabled the development of a robust multi-parametric assay for the identification of genes involved in DNA Damage Repair (DDR) in HeLa cells. The screen and follow-up experiments led to the discovery that the ubiquitin ligase E3 ubiquitin-protein ligase, Ring Finger Protein 8 (RNF8) is a central mediator in the DNA-damage response (Kolas et al. 2007). We further discovered how the ubiquitin ligase E3 ubiquitin-protein ligase, Ring Finger Protein 168 (RNF168) causes the cellular and developmental phenotypes characteristic for the Radiosensitivity, Immunodeficiency, Dysmorphic, Difficulties LEarning (RIDDLE) syndrome (Stewart, Panier, et al. 2009). I was also responsible for compiling a list of uncharacterised Open Reading Frame (ORF)s to identify novel DNA repair associated genes. From this list we implicated the MMS22L-TONSL complex in Double Strand Break (DSB) repair and described its role in the recombination-dependent repair of stalled or collapsed replication forks (O'Donnell et al. 2010).

Methods developed for the HCS on DSB repair were applied to another study focused on the discovery of genes involved in *VACcinia Virus* (VACV) infection (Beard et al. 2014). A brief summary of the challenges and my contributions to this study are described in section 2.2.

Parameter	Setting
Nuclei Detection Algorithm	B
Threshold Adjustment	1.5
Minimum Nuclei Distance	7
Individual Threshold Adjustment	0.4
Minimum Nuclear Area	70
Minimum Nuclear Contrast	0.05
Minimum Intensity Threshold per object	100

Table 2.1: Nuclei detection settings used in Acapella

2.1 A HCS screen to identify genes involved in DDR

2.1.1 Data acquisition

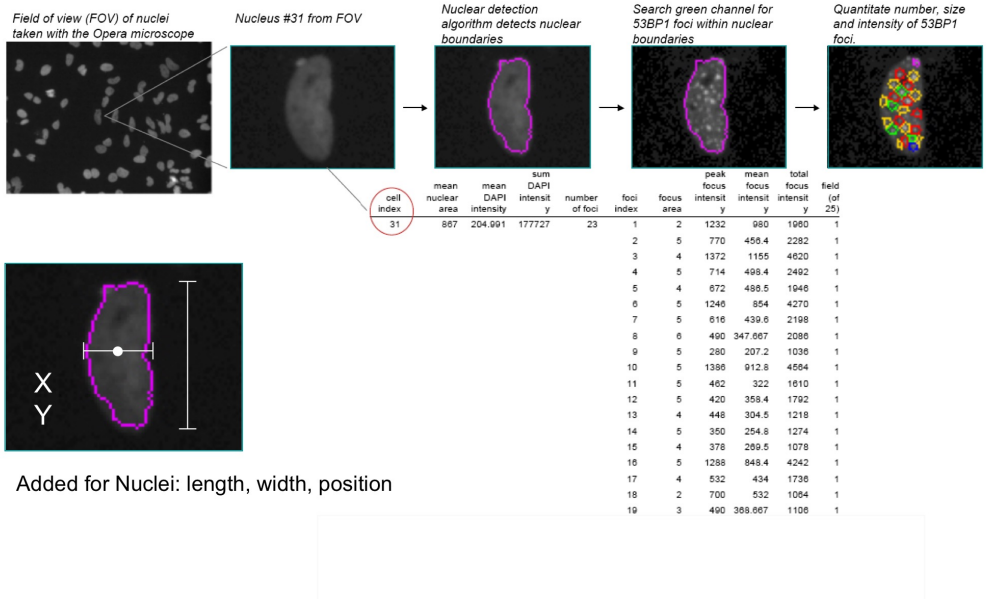
HCS data is derived from images using segmentation algorithms and markers that stain prominent cellular features to locate objects in microscopy images. These markers are usually fluorophores or fluorescent proteins that absorb and emit light at different wavelengths after light excitation.The blue fluorescent 4',6-diamidino-2-phenylindole (DAPI) is a commonly used marker for imaging and detecting nuclei. It intercalates into DNA and therefore allows visualisation of the nucleus in interphase cells. In this work DAPI stained nuclei were visualised with UltraViolet (UV) excitation on the Opera channel 1. A second marker, a fluorescently labeled antibody against Tumor Protein p53 Binding Protein 1 (TP53BP1) was detected in the second Opera channel with a green laser, used as a marker for DSB repair sites.

To analyse the images, we used the Acapella software developed by Evotec.

Parameter	Setting
Spot Detection Algorithm	A
Spot Minimum Distance	2
Spot Reference Radius	3
Spot Peak Radius	0
Spot Minimum Contrast	0.6
Spot Minimum To Cell Intensity	1

Table 2.2: Spot detection settings used in Acapella

Acapella provides a framework to access a wide range of image manipulation routines to extract features found in microscopy images. It is also possible to implement new analysis routines. Some of the standard scripts measure the intensity and area of nuclei and cytoplasm of a cell given that appropriate fluorescent markers are used. These scripts usually return a well average for each chosen feature. The script for this project required the addition of a spot detection algorithm within the nucleic area, the export of all measurements into text files for post processing in R and the export of image data to a Server Message Block (SMB) drive. After I developed the script, I optimised parameters to improve the object detection. For any image analysis script, parameter ranges had to be defined either as constants or dynamical parameters based on signal intensity and sharpness of objects due to focal point variations. Acapella provides a set of basic functions that are capable of detecting a wide range of nuclei and spot-like objects. After optimisation and testing of different settings on a reference set of images, we achieved the best performance with settings listed in Table 2.1 for DAPI stained Henrietta Lacks cervical cancer (HeLa) cells. For TP53BP1 spot detection, Acapella's spot detection segmentation was used with the settings shown in Table 2.2. The modified Acapella script provided the index parameters plate, well and image field together with derived parameters nucleus and foci number and nucleic coordinates x and y per image field. Further, the script



cell index	mean nuclear area	mean DAPI intensity	sum DAPI intensity	number of foci	foci index	focus area	peak focus intensity	mean focus intensity	total focus intensity	field (of 25)
31	867	204.991	177727	23	1	2	1232	980	1960	1
					2	5	770	456.4	2282	1
					3	4	1372	1155	4620	1
					4	5	714	498.4	2492	1
					5	4	672	486.5	1946	1
					6	5	1246	854	4270	1
					7	5	616	439.6	2198	1
					8	6	490	347.667	2086	1
					9	5	280	207.2	1036	1
					10	5	1386	912.8	4564	1
					11	5	462	322	1610	1
					12	5	420	358.4	1792	1
					13	4	448	304.5	1218	1
					14	5	350	254.8	1274	1
					15	4	378	269.5	1078	1
					16	5	1288	848.4	4242	1
					17	4	532	434	1736	1
					18	2	700	532	1064	1
					19	3	490	368.667	1106	1

Figure 2.1: Data capture from TP53BP1 and DAPI channels in DNA double strand break study. Nucleus area was detected using the standard procedures in Acapella. Data output is written as a cell and foci index file for further processing.

extracted the nucleic area, DAPI intensity, DAPI peak intensity, nucleus length and width as indicated by DAPI staining, foci intensity, foci peak intensity and foci area. All parameters were exported to text files after all fields in a well were processed. Figure 2.1 visualises the data extraction workflow. Based on comprehensive data analysis and comparison of all parameters I focused on the following three parameters: Total Spot Area (TSA), Total spot intensity (TSI) and Total Spot Number (TSN).

2.1.2 Data storage and management

When the DDR screen was initiated, few data storage platforms were available and most were not able to support very large and complex data sets typically obtained with HCS. This dataset contained more than 20,000 images with a total of 3×10^8 individual cells. Given that data handling in Excel spread sheets was limited to 65,000 rows at the time, I built a MySQL relational database to store all screening information and established a Samba storage server to manage the images. The Entity-Relationship (ER) model for the core part of the database used for large screens is shown in Figure 2.2. To allow easy navigation of the data and include cross references to existing meta information from other sources I built a user interface in Hypertext Preprocessor (PHP), for my collaborators. The web portal linked several databases (Search Tool for Recurring Instances of Neighbouring Genes (STRING), Information Hyperlinked Over Proteins (IHOP), Atlas, EnsEMBL, National Center for Biotechnology Information (NCBI) and Online Mendelian Inheritance in Man (OMIM)) to our RNA interference (RNAi) screen data sets for follow-up. I integrated R (Ihaka and R Gentleman 1996; Ihaka and Robert Gentleman 2012) and interactive Scalable Vector Graphics (SVGs) to allow interactive visualisation of the screening data and browsing of microscopy images (see Figure 2.3). At that time, the interactive tooltip SVGs were quite innovative, and are now routinely implemented in Data-Driven Documents (D3) (Bostock, Ogievetsky, and Heer 2011) using Hyper Text Markup Language 5 (HTML5). Commercial software packages to support meta analysis of initial hits such as Ingenuity (Qiagen) and Spotfire (PerkinElmer) existed but were expensive for academic labs and had focused specialisations, either pathway discovery or providing interactive tools to process and analyse raw data. Notably, the Open Microscopy Environment (OME) developed by Swedlow and colleagues (Swedlow et al., 2003) began as an open source academic software;

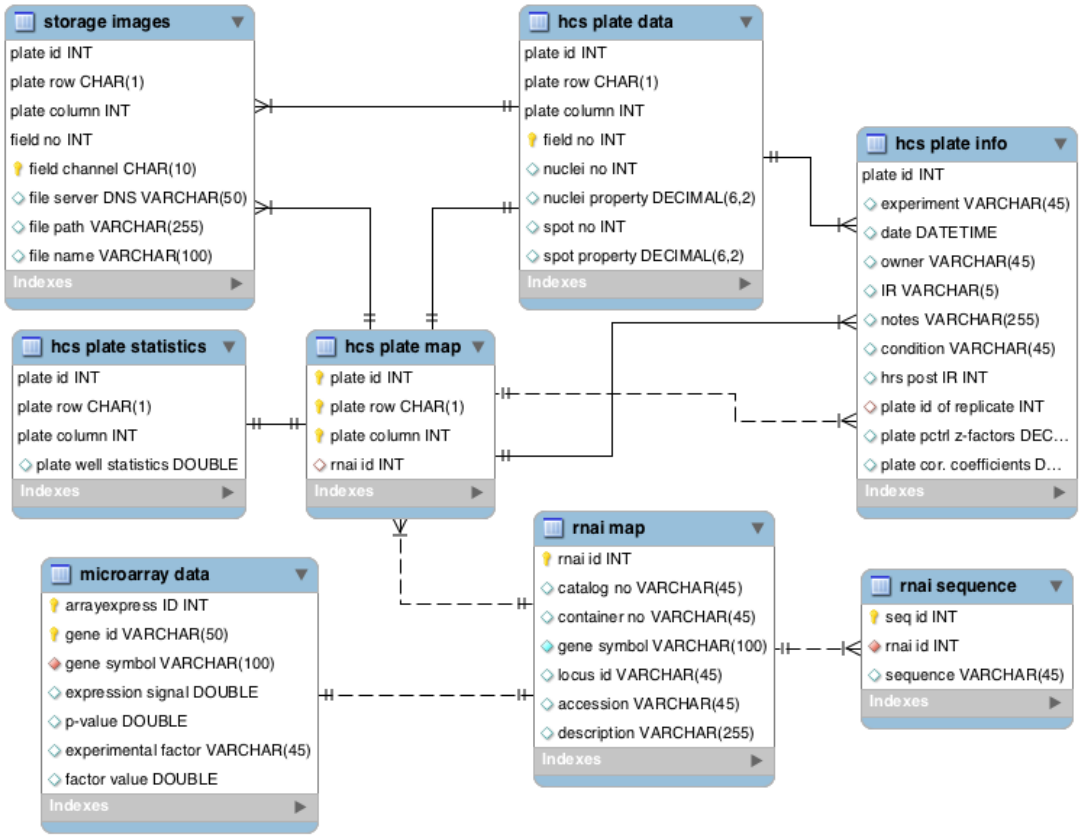


Figure 2.2: ER model used for storage of HCS data. Image segmentation parameters are added to 'plate data', additional plate quality control statistics are stored in 'plate info'. Summary statistics for each well for all derived properties are stored in 'plate statistics' to capture differences in variation, counts, quantiles, medians, means, modes, correlations within each well.

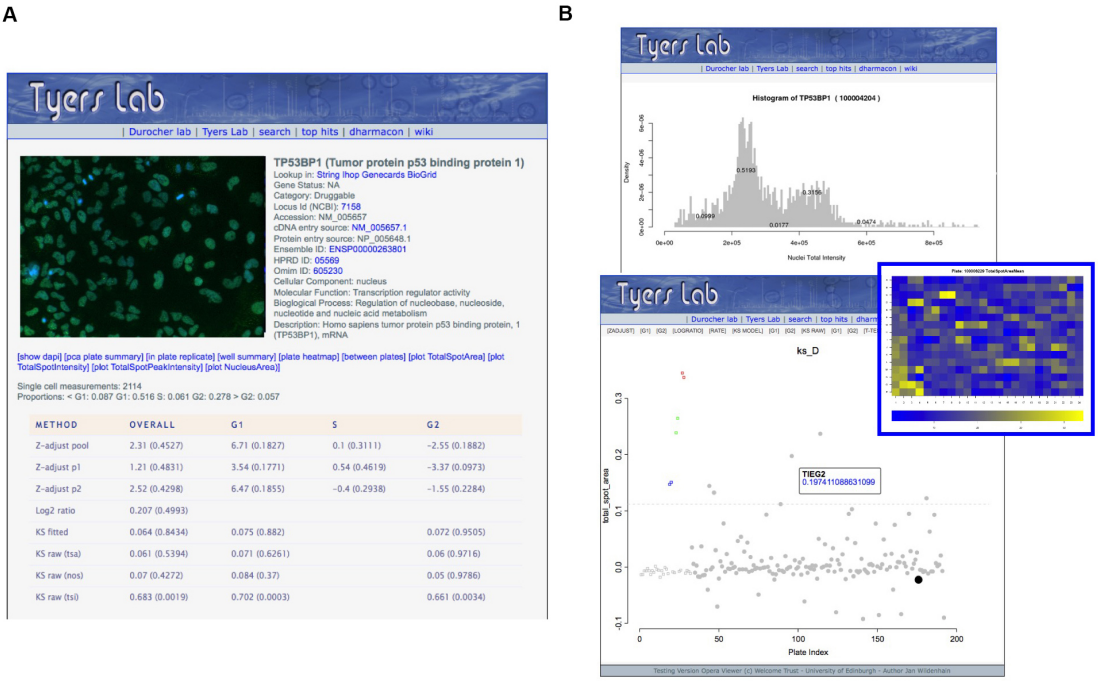


Figure 2.3: Properties derived from image segmentation parameters. **A** Summary of data statistics derived from high content data. **B** Visualisation of a DAPI distribution and a scatterplot of total spot area across a plate.

OME has since been elaborated into the OMERO platform[1] that can now be purchased from PerkinElmer as part of the web-based analysis suite Columbus. For specific types of screens, programming expertise is usually required to build or customise commercial data management frameworks. Various application notes and software packages for High Throughput Screen (HTS)/HCS data analysis have been published in the last few years, including cellHTS (Pelz, Gilsdorf, and Boutros 2010), HTSCorrector (Makarenkov et al. 2006), RNAither (Rieber et al. 2009), HCS-Analyser (Ogier and Dorval 2012) and EBImage (Pau et al. 2010). These packages are usually tailored to institutional workflows and support a small user base. More successful image analysis suites, such as Cellprofiler (Anne E Carpenter et al. 2006), ImageJ (Schneider, Rasband, and Eliceiri 2012) and Fiji (Schindelin et al. 2012), are open source to engage users in future development. With its many challenges and contributors, the field of image-based analysis is now sometimes referred to as Bioimage informatics (Eliceiri et al. 2012). My HCS image processing and data analysis procedures are available as Acapella and R scripts on GitHub[2] or as shiny web applications. Software code and applications are available on sysbiolab.bio.ed.ac.uk (Wildenhain, Fitzgerald, and Tyers 2012), chemgrid.org (Wildenhain, Spitzer, Bellows, et al. 2015) and boxplot.tyerslab.com (Spitzer, Wildenhain, et al. 2014).

2.1.3 Data normalisation

A challenge in designing large scale experiments is to collect data without bias. In screening, plate corner effects and library design influence the hit selection. A lot of work has been done by Terry Speed and others in addressing biases in microarray design (Lönnstedt and T Speed 2002; Smyth, Yang, and Terry Speed 2003; Quackenbush 2002). Spotting of repeats on arrays, the number of repeats

[1] https://www.openmicroscopy.org
[2] https://github.com/jwildenhain

and changing the labelling direction (dye swaps) have been found to be important to apply robust statistical methods. To my knowledge, in HTS such high level recommendations have not been established. It has been suggested that molecules or RNAi's be randomised on plates before screening, however, this is often not practical. Dharmacon and Qiagen provide their libraries in alphabetical order. Similarly, libraries provided by chemical vendors, like Maybridge, are sorted by the enumerated initials of the synthetic chemist so that similar compounds are grouped together. Technical and biological repeats are important for robust data analysis. For this study, we designed the experiment with a technical and a biological repeat as shown in Figure 2.4. HCS data analysis methods rely heavily on positive and negative controls to assess data quality and significance of effects caused by the applied chemicals or conditions. Since control wells are usually pipetted with separate instruments and reagents, they are a source of additional independent biases. This can lead to rejection of large parts of the data and can be avoided by introducing a cross-correlation analysis of the experimental data. A common choice to test for screen quality is the Z-factor (J.-H. Zhang, Chung, and Oldenburg 1999) or Strictly Standardised Mean Difference (SSMD) (X. D. Zhang 2007) value, both methods rely on the control groups and therefore expose an Achilles heel if controls and samples are handled differently or controls were chosen poorly. The Z-factors for two positive controls, Proliferating Cell Nuclear Antigen (PCNA) and RADiation repair gene 51 (RAD51), illustrate this (Figures 2.5A and 2.5B). with different distances to the negative controls. First, the strength is different due to the effect of silencing on the cells. Second, the moderate Z-factors are present in different plates of the screen. A better method is a Pearson correlation Quality Control (QC) plot. It requires at minimum either a technical or biological replicate and a set of non-replicates that have been created in the same batch. Figure 2.5C illustrates the coefficients for the replicates and non replicates and they show very distinct clusters between the average coefficients for the biological repeats (Q1/Q3 and Q2/Q4) and the background (Q1/Q2,

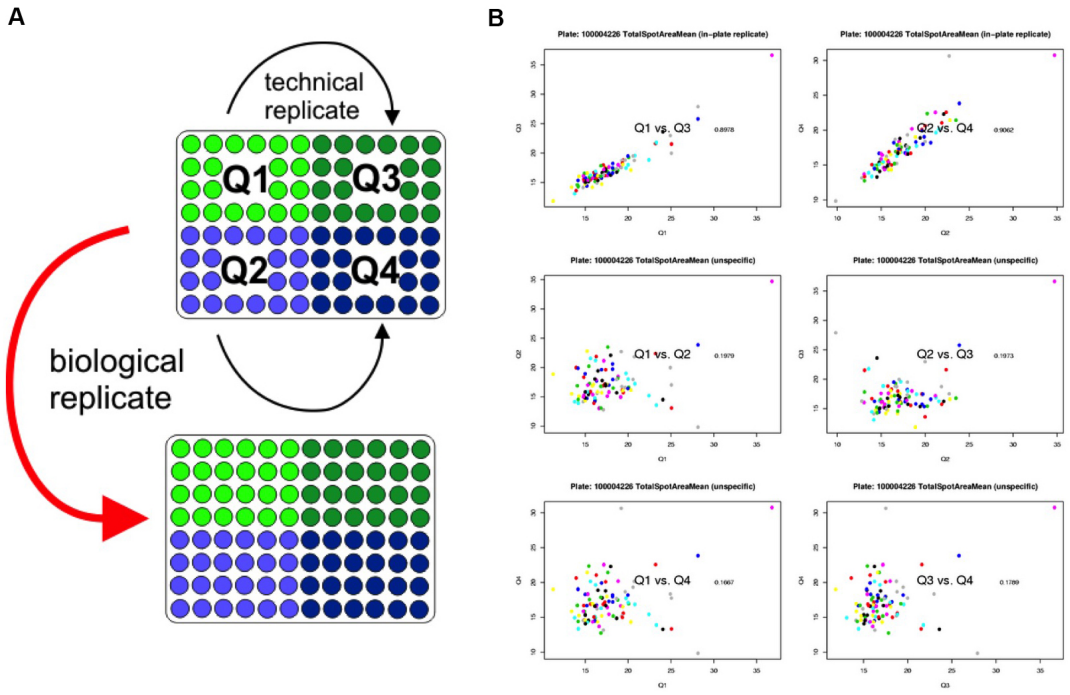


Figure 2.4: Experimental design for DSB HCS study. **A** 384 well plates are deconvoluted into 4 quadrants (Q1-Q4) with 2 repeats of 80 siRNAs. A second plate with the same layout has been screened as a biological replicate. For each well, 25 images (fields) were taken. **B** Scatterplot for one property between all quadrant combinations. The two technical replicates are shown in the top two scatterplots, the remaining scatterplots are the background versus non-replicates.

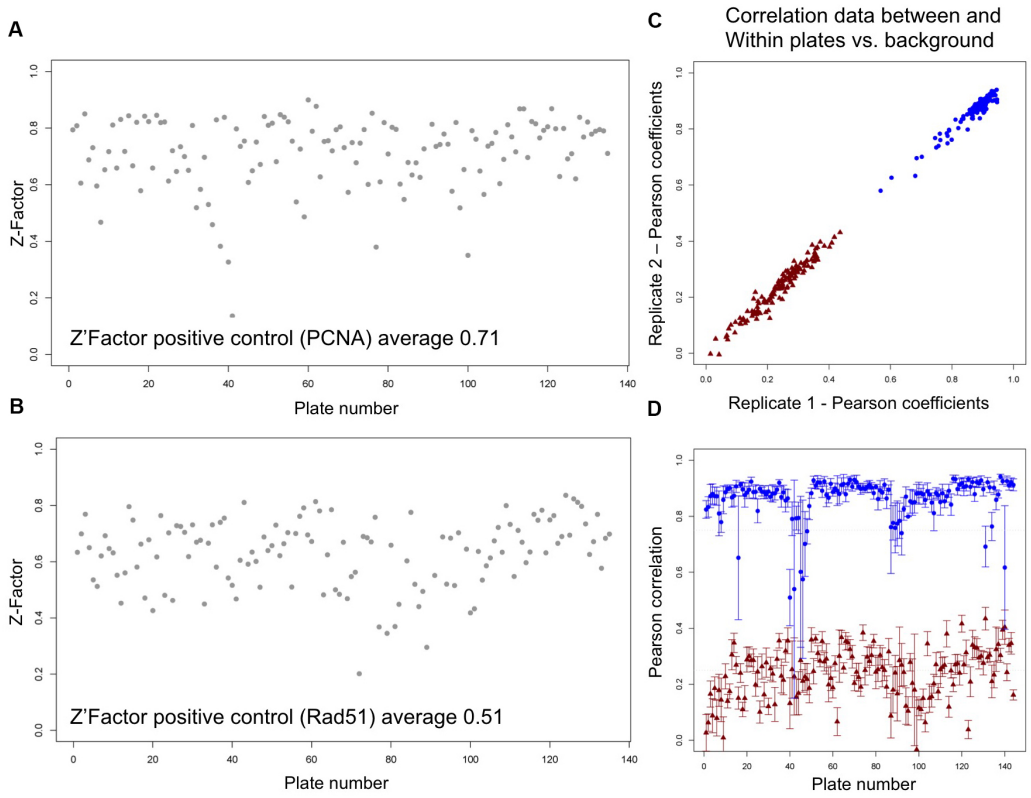


Figure 2.5: QC for HCS/HTS. **A** Z-Factor for positive control PCNA across plates. **B** Z-Factor for positive control RAD51 across plates. **C** Pearson correlation scatterplot between replicate (blue) and non-replicate (brick-red) plate pairs. **D** Pearson correlation plate statistic per plate. Error bars show SD between replicates.

Q2/Q3, Q1/Q4, and Q3/Q4) of the two replicates. Using the average Pearson coefficients and Standard Deviation (SD), we can see in Figure2.5D that for the plates 39 to 44 the TSA signal quality drops and error bars overlap with the background estimate. Similarly, the plate Pearson correlation QC plot highlights the outlier plates 16 and 139 that would not be noticed using a Z-factor or SSMD.

A broad range of factors can influence the experimental outcome and lead to screening bias. Systematic errors can sometimes be controlled by optimising the protocol. Other noise might depend on temporal factors and increase over

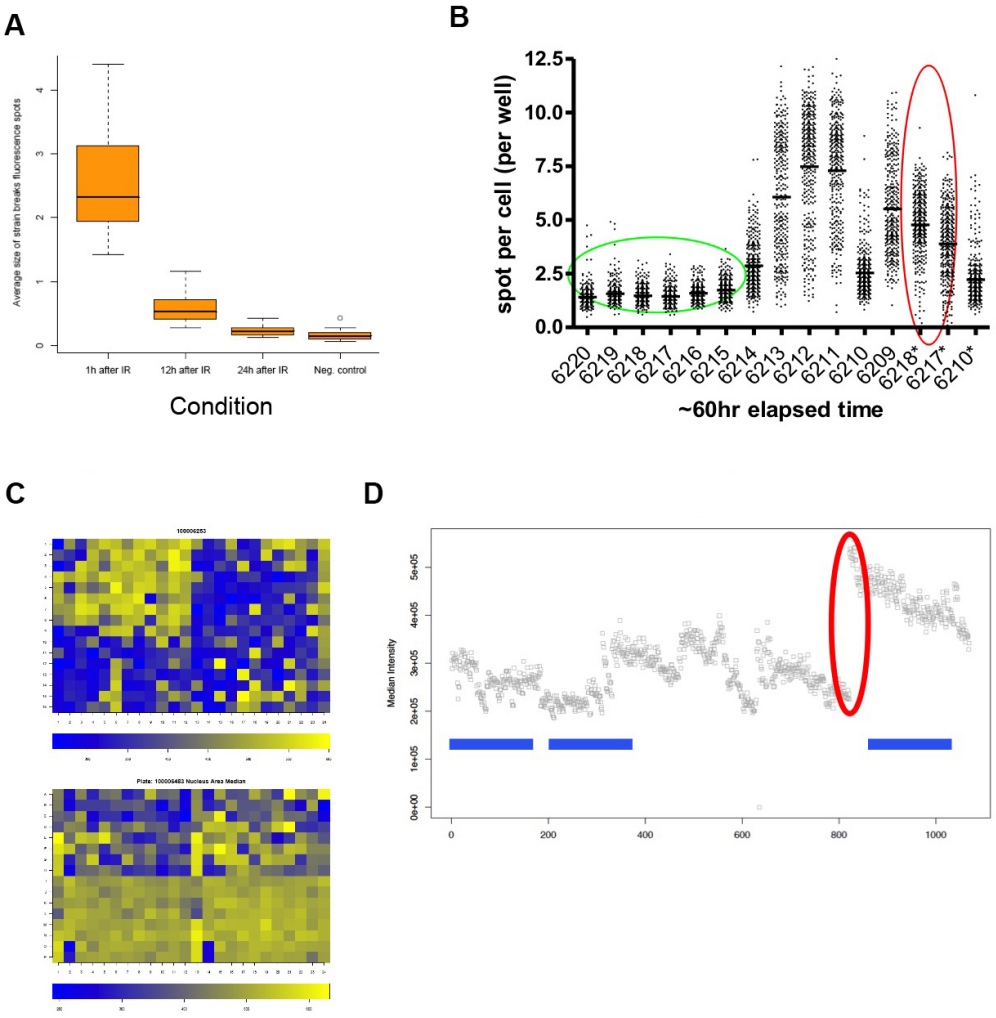


Figure 2.6: Data capture bias in DNA DSB study. Signal variation shown for **A** Number of TP53BP1 foci after 1, 12 and 24 hours after UV irradiation **B** Number of TP53BP1 foci increases over time due to anti-body precipitation: initial read 24h after sample preparation (green); second read of plates 6217 and 6218 after an additional 60h incubation (red) **C** Deconvolution of 384 well plate based on 96 well head used for transfer reveals spatial variation within the plate due to liquid handling robot (top: DAPI intensity signal, bottom: HeLa cell count) **D** Differences in DAPI median intensity between screening batches due to variations in DAPI concentration (blue boxes) and UV source replacement (red oval).

time. An example is the exposure to light or radiation. Similarly, the stability of chemical and biological processes can influence outcome. With increasing complexity of experimental designs the number of factors that can affect the outcome increases. In large HTS studies robust methods are needed to allow comparison of the data across different experiments. A few examples of factors that influence HCS readouts are illustrated in Figure 2.6. The sample position in 96, 384 or larger well plates can influence the assay results. These effects may be due to a combination of factors and there are several methods that can be employed to correct these effects (Dragiev, Nadon, and Makarenkov 2012; Malo et al. 2006; Birmingham et al. 2009). First, the data variation is commonly used to weigh experimental values. A typical procedure is the standard score, where a measure of central tendency for an experiment (usually mean or median) is subtracted from each value of an experiment followed by division by a measure of variation. This method is simple and robust for normally distributed data to identify significant outliers with respect to the variation of the population. Secondly, the application of a normalisation method can reduce a data inconsistency bias if it exists. A common method is median normalisation or quantile normalisation to calibrate values between different wells and plates within and between studies. Third, increasing the sample size through experimental repeats can improve and compensate for inconsistent outliers; a highly valued method is ANalysis Of VAriance (ANOVA) in this regard. For ANOVA several repeated measures are important to increase statistical power. Experimental repeats in large scale studies are a particular concern as they raise the research cost, limiting the applicability. A detailed knowledge of the experimental workflow can guide the selection of normalisation procedures to remove spatial effects in the data. Figure 2.7A shows an example of low cell numbers in the lower left corner of a 384 well plate. Since such effects are only locally present a spline fit or a regression model can compensate for such outliers without the need to discard this plate area or heavily weigh putative outliers. The difference in cell counts in

the first two columns and the remaining columns can be removed with a median quartile normalisation between control and experiment. The intensity levels of phalloidin, a selective marker of F-actin in fixed cells, show a pattern of column intensity shifts. This can happen in robotic liquid delivery with an eight-span pipetting head into a 96- or 384-well plate. A different parameter, the median virus intensity, is less affected by this column effect. The corresponding heatmap is shown on the right in Figure 2.7A. These systematic effects can be addressed with a combination of regression smoothing and median quartile normalisation (bottom row in Figure 2.7A). Another example of LOcally WEighted Scatterplot Smoothing (LOWESS) based normalisation is shown in Figure 2.7B. Finally, to aggregate the data for each well, I calculated adjusted Z-scores for different parameters with the following formula

$$adjusted\ Z = \frac{\mu(x) - \mu(X_{ctrl})}{\sigma(X_{ctrl})}$$

where x represents all single cell measurements of the cell population for a siRNA and X_{ctrl} the negative control population measurements (well averages). The corresponding p-value was calculated using the Kolmogorov-Smirnov distance between x and x_{ctrl}. For every plate, a beta distribution is fitted to the Kolmogoroff-Smirnov (KS) data to derive the corresponding p-values. The adjusted Z-score incorporates some degree of variation of the single cell measurements in relation to the variation within the negative controls on the plate. All adjusted Z-scores of the DMSO negative controls are within -4 and 4 σ from $Z = 0$. The control distribution can therefore be described as a normal distribution with $\mathcal{N}(0, 1)$. We reported the adjusted Z-score for TSA as the preferred property to segregate positive and negative controls (Kolas et al. 2007), similarly a score was calculated for TSI and TSN for comparison.

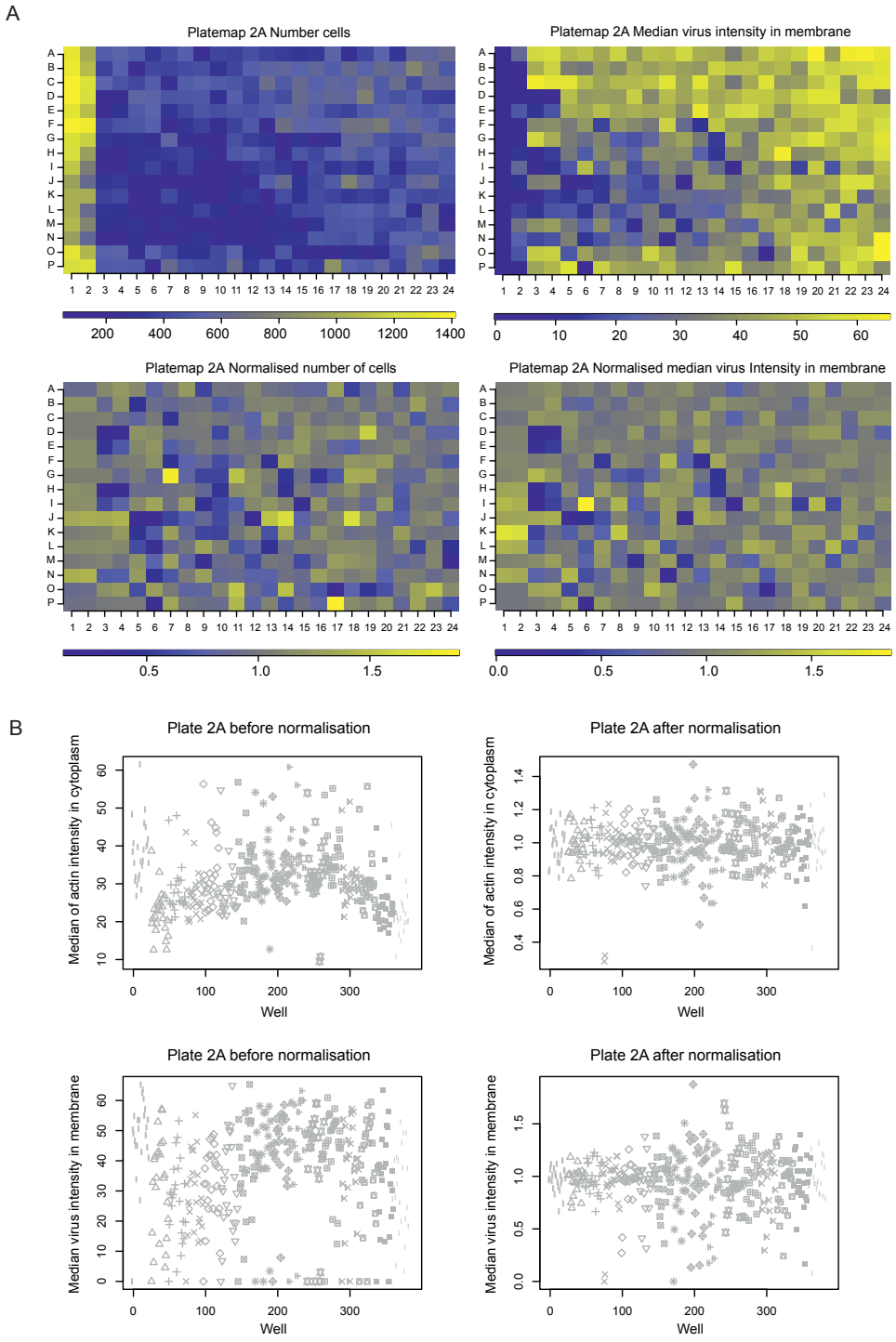


Figure 2.7: Visualising plate spatial effects. **A** Plate heatmaps before and after normalisation of nucleus counts based on segmentation and median virus intensity. **B** Parameters for actin median signal intensity in cytoplasm and virus median intensity sorted by plate, row and column (symbols represent rows A to P).

2.1.4 Data refinement

Another important issue in HCS is the sample size which translates into the number of images that are acquired for each well for robust data analysis. When working with RNAis or small molecules, toxicity effects can reduce the number of cells in each well. The minimum number of images that should be taken for each well for effective sample sizes, can be determined using permutations (Odén and Wedel 1975). First, define the test hypotheses

$$H_0 : \gamma_1 = \gamma_2$$
$$H_1 : \gamma_1 \neq \gamma_2$$

and select a test for paired difference (here KS test) with a distance measure D between two samples with corresponding empirical Cumulative Density Function (CDF)

$$D = \sqrt{\frac{mn}{m+n}} \sup_x |\Gamma_1(x) - \Gamma_2(x)|$$

The level of significance (e.g. $\alpha = 0.05$) defines the outcome

$$\delta = \begin{cases} H_0, D \leq c \\ H_1, D > c \end{cases}$$

where the threshold c depends on the level of significance α and can be derived from the condition

$$\alpha = \mathbb{P}(\delta \neq H_0 | H_0) = \mathbb{P}(D \geq c | H_0)$$

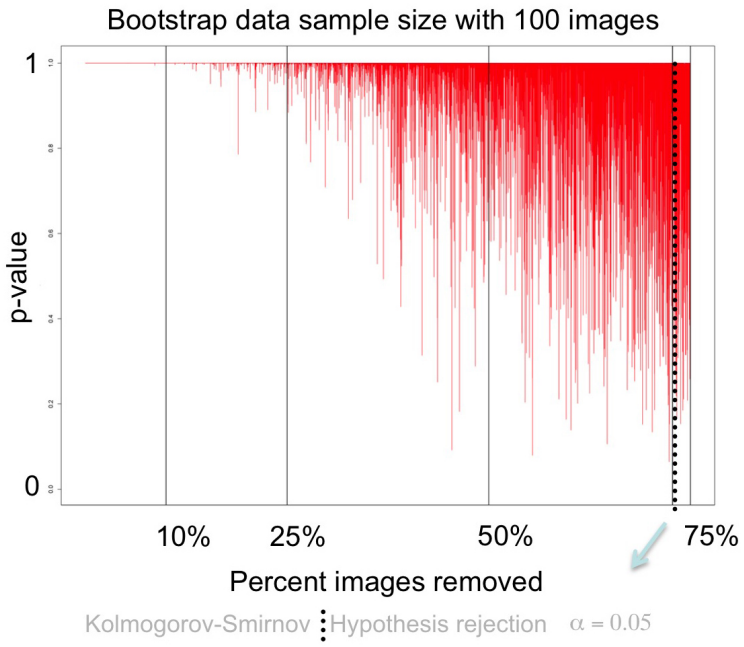


Figure 2.8: Permutation testing for hypothesis rejection for a non-targeting siRNA. In this bootstrap sample a rejection of H_0 at $\alpha = 0.05$ occurs after truncating 71 images.

In this context, the margin of error for an approximation of the p-value is proportional to the number of permutations
$\binom{n}{k} = {}^nC_k = \frac{n!}{k!(n-k)!}$. Assume that we have $n = 200$ images and we assign $k = 100$ images randomly to γ_1 and the remaining images to γ_2. Since there are many possible permutations, the p-value needs to be approximated by testing a reasonable subset of the possible permutations. Figure 2.8 shows the results of a permutation analysis for non-targeting siRNA with $n = 200$ and $k = n/2$ where n is iteratively reduced every 500 permutations to find the minimum number of images required to not falsely reject H_0 due to a p-value < 0.05. When $n = 58$, we are likely to falsely reject H_0. This means that in our assay, at least 30 images are required per condition for robust data analysis. For the screen we chose to take 50 images per siRNA for each biological replicate to account for reduced cell populations due to toxicity of the siRNAs. Since the parameters derived from the images

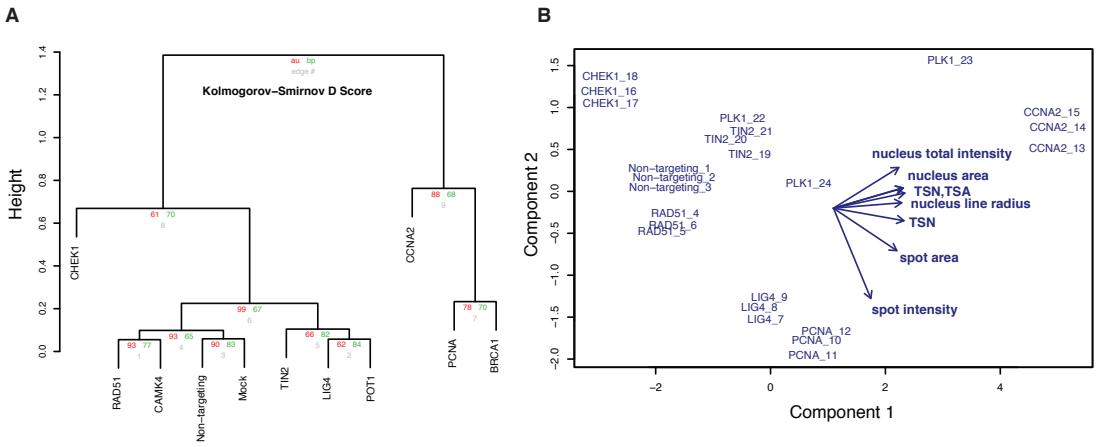


Figure 2.9: Distance comparison between different siRNA with KS Hierarchical Clustering and PCA. **A** Average hierarchical clustering of different RNAi based on TSA KS distance. **B** Biplot of the first two principal components with different RNAi based on the well averages are shown. Arrows indicate the directionality of the parameter eigenvectors.

followed different distributions I chose one parametric and two non-parametric statistical tests to assess differences between the samples. Welch's t-test (Welch 1947), KS test (Kolmogoroff 1941; Smirnov 1948) and Wilcoxon signed-rank test (Wilcoxon 1945) for independent samples were compared using average linkage hierarchical clustering (McQuitty 1960) and tested for correct clade assignment. All three tests performed well while the KS test was the most sensitive (example shown in Figure 2.9). For one parameter, nucleus size, the t-test showed the best performance because the sample distribution resembled a Gaussian distribution.

The large number of data points for each RNAi allowed me to generate a Fluorescence-Activated Cell Sorting (FACS)-like DNA content profiles for each RNAi probe based on the DAPI signals. Cells with extremely low or high DAPI intensity values were removed because they correspond to cells with less than $1N$ or more than $2N$ DNA contents. As shown in Figure 2.10, this procedure

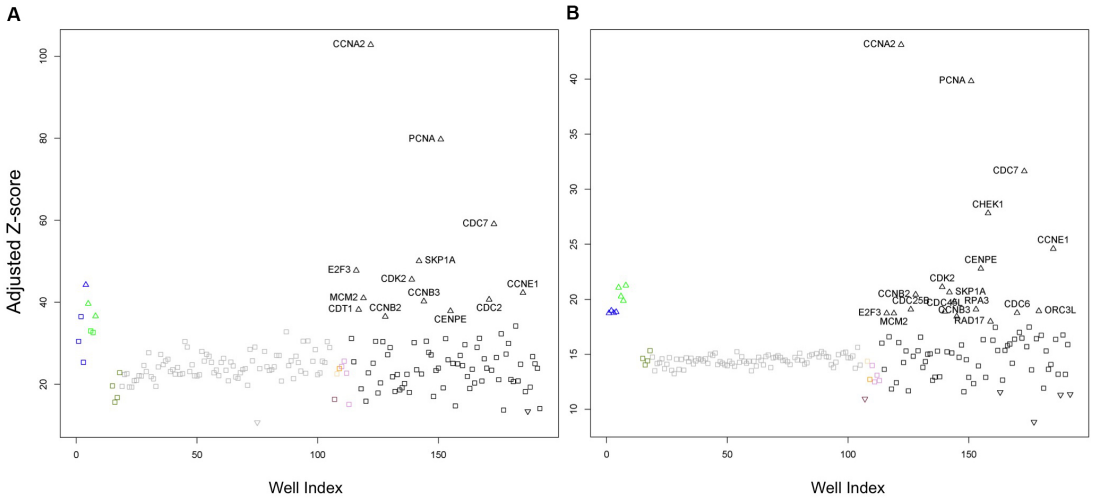


Figure 2.10: Single cell filter to reduce noise in HCS data. **A** Adjusted TSA Z-Score for a plate before single cell filtering. **B** Same plate after filtering. The filter removed cells without TP53BP1 foci, very low DAPI content or more than 2N DNA content based on DAPI.

greatly reduced the variance in the negative control, positive control and sample wells. The variation within the negative and positive controls is reduced and this improves the identification of weak hits. Further, DAPI intensity enabled the inference of potential effects on the cell cycle without using an additional fluorescent marker. Figure 2.11 demonstrates how the targeted silencing of a cell cycle checkpoint gene increases the number of cells that show signs of aneuploid DNA content.

The ability to segregate cells based on DNA content allowed me to build a model to detect the modes of a bimodal distribution to assign cells to five groups with DNA content below Growth phase 1 (G_1), G_1, S phase, Growth phase 2 (G_2) and above G_2 DNA content. Given the data $F(x)$, the model is defined by

$$F(x) = pF_1(x) + (1-p)F_2(x)$$

and we assume that the two distributions F_1 and F_2 representing 1N and 2N DNA

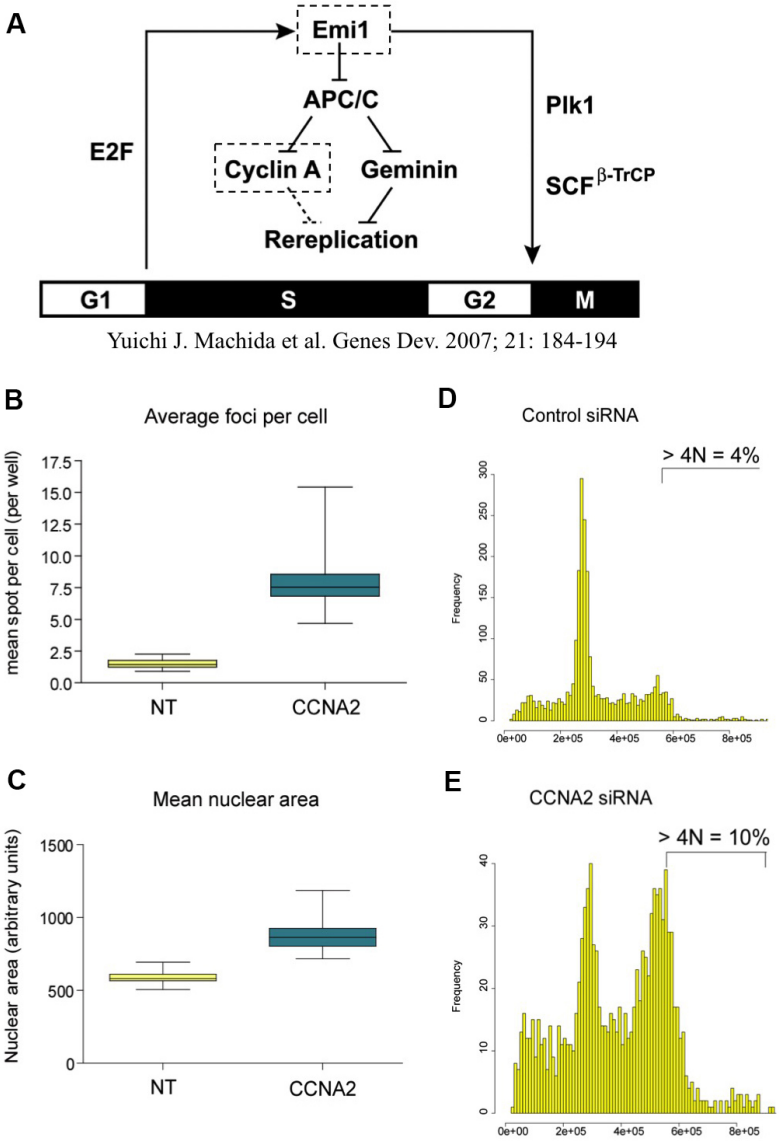


Figure 2.11: Example of cell cycle inhibition using RNAi. **A** Inhibitory regulation of re-replication by cyclin A (Machida and Dutta 2007). Effect of inhibition of cyclin A (CCNA2) on TP53BP1 foci formation (**B**) and nucleus area (**C**) compared to negative control. DAPI intensity profile for negative control (non-targeting RNAi) (**D**) and for CCNA2 (**E**).

contents are Gaussian. The aim is to estimate the parameters p, σ_1, σ_2, μ_1 and μ_2 (see Algorithm 1 for details). The algorithm bins the data and tries to identify the G_1 peak. If a single high intensity peak is identified, the mode is calculated from $F(x)$. Similarly, the mean μ and the variance σ are estimated from the data adjacent to the mode for a bin width of ± 4. Based on the parameters, a sampling distribution Φ is bootstrapped from $F(x)$ to calculate the models corresponding to the Akaike Information Criterion (AIC). This procedure is repeated for $Term$ iterations. The parameters from the best fitted model (p_{opt}) are used to represent the DNA content estimates. The bootstrapped distribution Φ of the best fit is removed from $F(x)$ before the same algorithm is applied to identify the G2 peak. Since some RNAis have strong effects on the DAPI measurements, the modes estimated from non-targeting controls are used as starting points for the RNAi sample wells. The adjusted Z-scores for the G_1 and G_2 phase correlate well (Figure 2.12). The spread of the values suggest that a high level of variation is present in the data with few outliers. The significance test was performed using the Mahalanobis Distance (MD) and hypothesis testing for $\alpha = 0.05$ (Mahalanobis 1936).

2.1.5 Screen validation with siRNA from other sources

After the initial screen the hit calling adjusted by the large pool of non-targeting siRNA controls suggested an unexpectedly high off-target rate of the Dharmacon library. Therefore we selected a subset of gene targets (476) for confirmatory experiments with two other siRNA libraries, Qiagen and an in-house endoribonuclease-prepared siRNA (esiRNA) library (Figure 2.13A). 175 of these genes were known to be involved in the different DNA repair pathways. These 476 siRNAs were screened following the same protocol for all three libraries. The other two libraries had a much lower hit rate than the Dharmacon library (Figure

Algorithm 1 Find cell cycle stage based on DNA content

Require: $F(x)$

repeat
  $B \leftarrow$ bin-dapi-profile(x)
  $\mu \leftarrow$ bin-frequency-median(B)
  $\nu \leftarrow$ bin-frequency-mad(B)
  $\beta \leftarrow 10$
  $\epsilon \leftarrow 0.0001$
  while $\sum \psi \neq 1 \vee Term$ **do**
    for all i in B **do**
      if $B(i) > \mu \wedge \psi(i+1) = 0 \wedge \psi(i-1) = 0$ **then**
        $\psi(i) \leftarrow 1$
      else
        $\psi(i) \leftarrow 0$
      end if
    end for
    $\mu \leftarrow \mu + (\nu\epsilon)$
  end while
  $\nu \leftarrow$ detect-mode(B, ψ)
  μ,$\sigma \leftarrow$ detect-parameters(B, ψ)
  $\Phi \leftarrow$ bootstrap-parameters-(x, μ, σ)
  $p \leftarrow \mathrm{AIC}(x, \Phi)$
  $\beta \leftarrow \beta + 1$
until $Term$
return p_{opt}, Φ_{opt}, $F'(\mu, \sigma)$

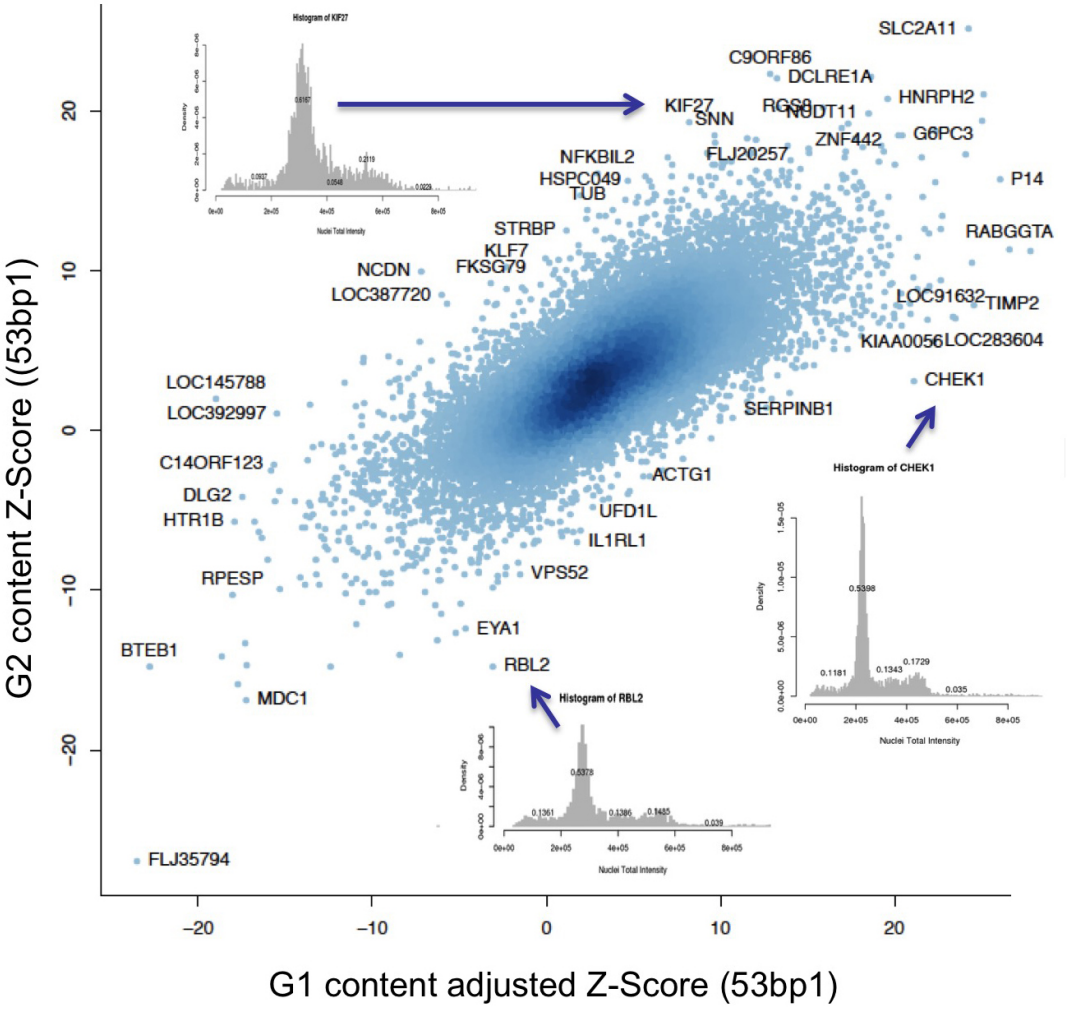


Figure 2.12: Comparison of G_1 and G_2 adjusted Z-Scores. The KS model produces a non-linear spread compared to the Z-score highlighting additional genes. The non-directionality of the KS method looses information of increased or decreased foci. The gene with the lowest adjusted Z-score is RNF168 (FLJ35794) and RNF8 is not shown (TSA adjusted Z-Score < -30.)

2.13B,C). Only 21 siRNAs were hits in all three libraries and 119 did not show an effect on TP53BP1 foci in any of the libraries (Fig. 2.13C). I calculated Gene Ontology (GO) enrichment p-values for all genes that were classified as hits based on their Z-adjust score in at least two of the three RNAi libraries (Figure 2.13D). Within this set of 153 genes we found 38 previously uncharacterised genes that would be interesting for follow-up studies to investigate their potential role in DNA damage repair or related functions. The workflow was highly robust and all technical replicates showed a correlation for each parameter greater 0.85 with a random background Pearson correlation below 0.2. One potential reason for the observed discrepancies might be the frequent off-target effects of imperfectly-matched 3 UTRs (Fedorov et al. 2006; Jackson and Linsley 2010). At the time, the siRNA sequences from Dharmacon were a trade secret and not revealed to users of the library. According to Dharmacon, each gene was targeted by a pooled set of 4 different siRNAs with a chemical modification designed to reduce off-target effects. A similar approach has been taken by Qiagen, whereas the esiRNA pools used (R. Kittler et al. 2007) were created by an in-house facility, leading to variable pools of up to 12 unique sequences per gene. The lack of transparency and limited information for each of the siRNA sources made it difficult to evaluate and compare the complete hit lists. Nevertheless, a reference dataset of genes where at least two siRNA sources confirmed significant changes in induced TP53BP1 foci was defined. This reference set was used to determine the cut-offs for the quantitative parameters to estimate false discovery rates. Due to limited resources for further experiments, these efforts were discontinued.

To assess the accuracy of normalised intensity and area of the foci, we tested how well the adjusted Z-scores would perform using the list of 175 curated genes as a gold standard gene set to estimate True Positives (TP) and False Negative (FN). We assessed the sensitivity $s = \frac{\text{TP}}{\text{TP+FN}}$ of intensity and area of TP53BP1 foci in identifying genes likely to be involved in DDR (Figure 2.14). Increasing the

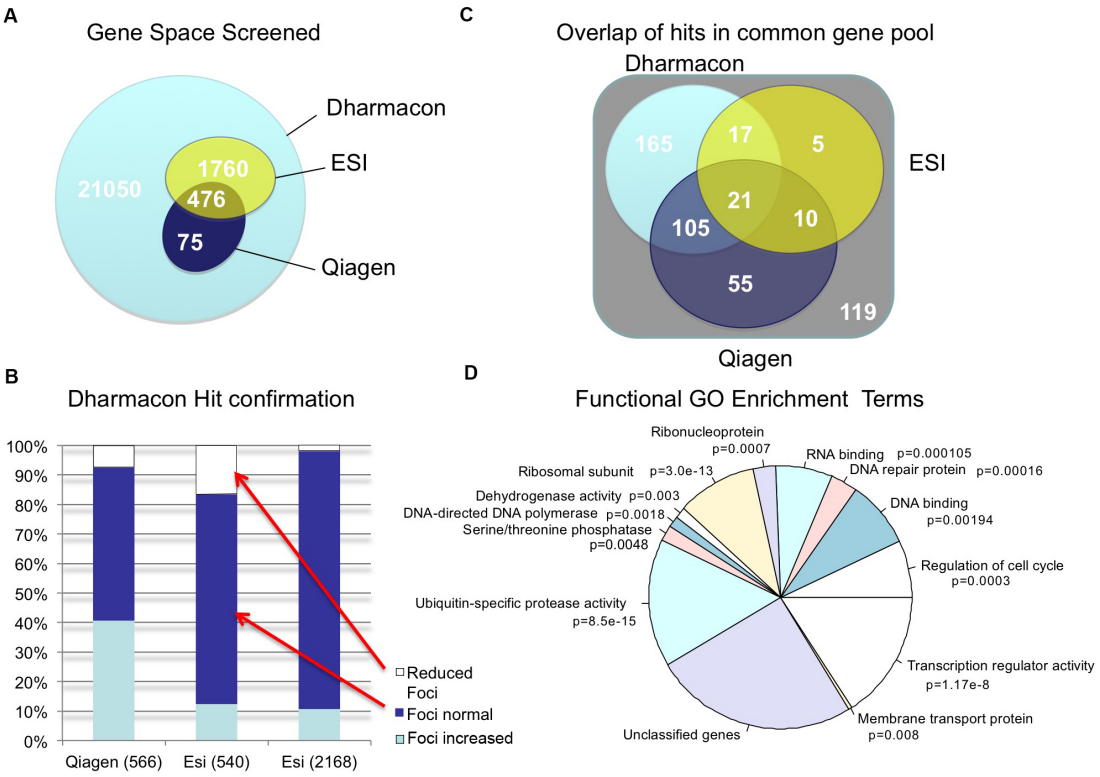


Figure 2.13: Robust DNA repair core gene set for feature evaluation and follow-up studies. **A** Venn diagram shows hit siRNA overlap between Dharmacon, ESI and Qiagen. **B** Effect on the number of TP53BP1 foci obtained with Qiagen and ESI (two biological repeats). Colours in bars indicate reduced, normal and increased number of foci. **C** Overlap in hits between siRNA pools from Dharmacon, ESI and Qiagen. **D** Functional GO enrichment of hits in the shared gene pool. P-values were obtained with a hypergeometric test using Bonferroni correction. The GO annotations were obtained from HPRD.

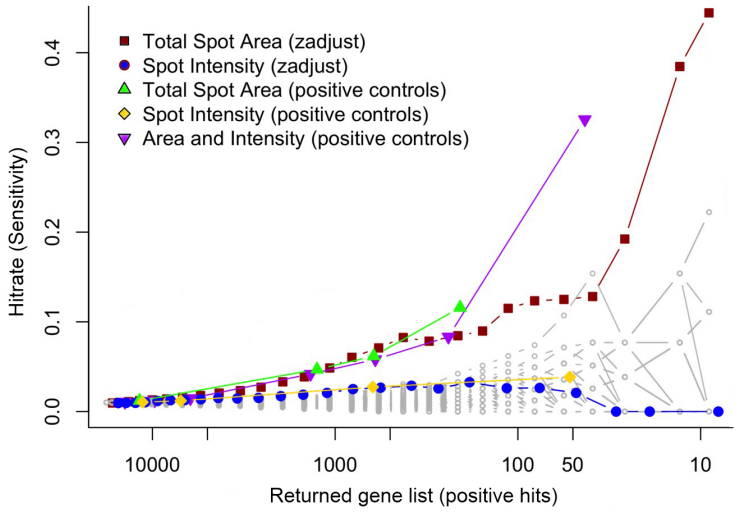


Figure 2.14: Comparison of the sensitivity between different parameters. Sensitivity calculation shows adjusted Z-Scores for TSA and TSI with increasing number of hits. The TP list consists of 175 genes that have been annotated to be involved in DNA repair. Grey lines represent hit rates from random sampling.

number of hit siRNAs in a step-wise manner, TSA clearly outperforms TSI in this context.

2.1.6 Assessing significance with the non-parametric KS test

The Z-adjust measures of the TP53BP1 foci represented the data well. However, in the Dharmacon DNA damage screen many siRNAs were classified as hits. To assign significance to the measurements, I explored different statistical tests and found the KS distance from each targeted gene to the non-targeting siRNA on each plate to work very well. After applying a kernel density estimation algorithm, the KS distances are represented by a beta distribution. For each plate, p-values were calculated based on kernel density estimation. A volcano plot of the adjusted Z-score of TSA shows the best separation between the negative and positive controls

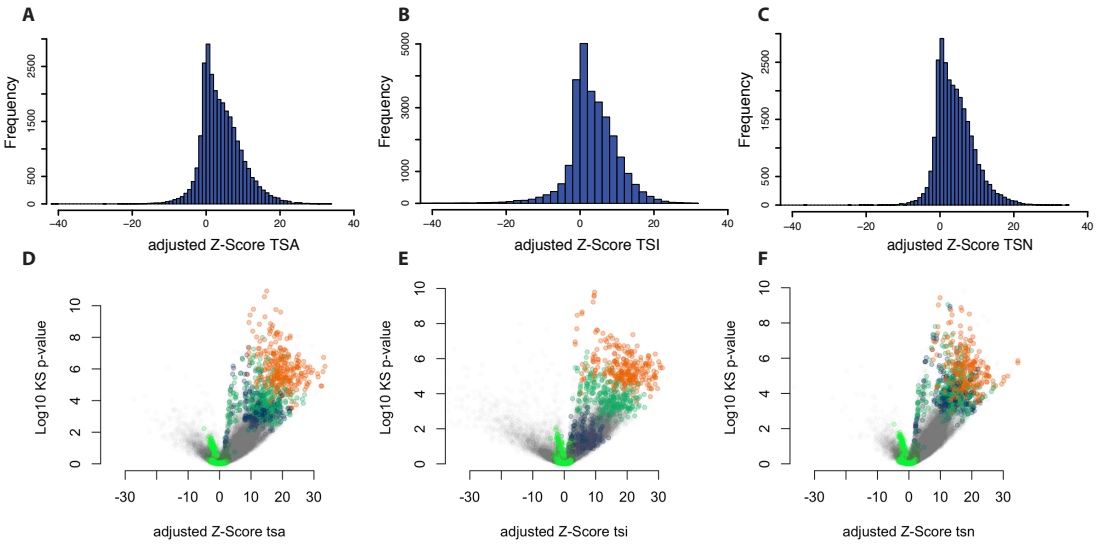


Figure 2.15: Comparison between TSA, TSI and TSN and respective KS p-values. Histogram of foci properties **A** TSA **B** TSI and **C** TSN with the corresponding volcano plot showing the KS p-values (below). The colours represent the controls non-targeting (bright green), RAD51 (blue), LIG4 (green) and PCNA (orange). The best separation between the different positive controls is achieved with TSI. The best separation between non-targeting control and positive controls is achieved with TSA.

while allowing to clearly distinguish the different positive controls. To illustrate, Figure 2.15 shows volcano plots for the different adjusted Z-Scores and their corresponding KS p-values.

Initially the reliance on counts in this study has been problematic as shown in Figure 2.6. With the improved imaging routine TSN increased it reliability. After introducing the KS based analysis I reanalysed the TSN data. Figure 2.16 shows a good correlation between TSA and TSN foci data. Residuals deviating from a linear model between both properties might have biological relevance. Noticeable are proteasome subunits with increased foci numbers. In comparison a correlation between the TSA and TSI produces an increased level of variability with decreased correlation (Figure 2.17). As expected, silencing the TP53BP1

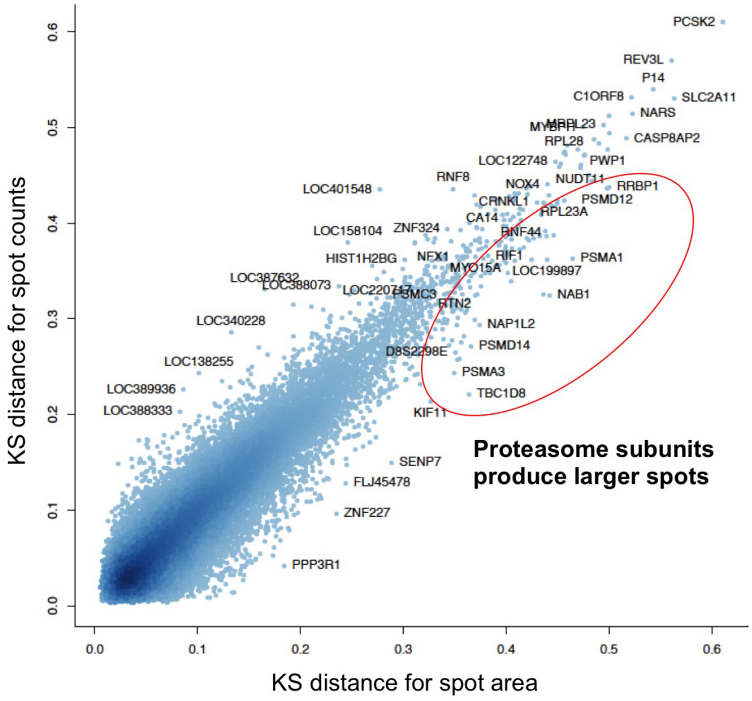


Figure 2.16: Comparison between TP53BP1 foci counts and area with a Pearson correlation coefficient of 0.94. The absolute distance from the negative control is shown. The signal for TP53BP1 is shown in top left corner.

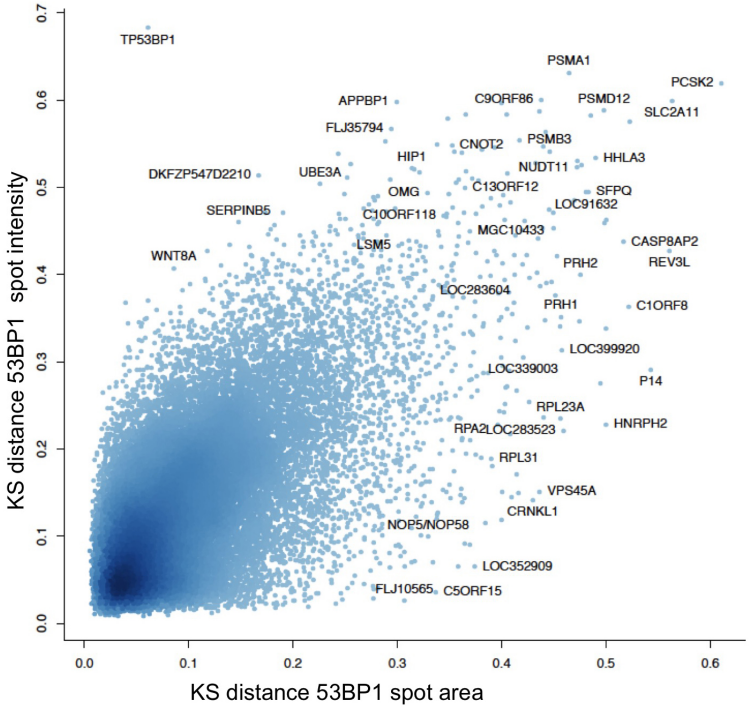


Figure 2.17: Comparison between foci intensity TSI and foci area TSA of TP53BP1 with a Pearson correlation of 0.68.

leads to the disappearance of foci (Figure 2.17). As demonstrated earlier, intensity is a less reliable measure based on the reference 'gold' standard and the KS based p-values used in this study. Intensity shows a correlation with the area which indicates that the variability might contain additional information that can be explored in a multivariate approach. Interestingly RNF8 and RNF168 are not obvious outliers using the KS distance from negative control. In contrast the t-test for the parameter nucleic size correlates tightly with the KS distance as shown in Figure 2.18A compared against the non-targeting siRNA control population. The performance between both tests is very similar since Nucleic Size Area (NSA) data was well represented by a normal distribution. The influence on the nucleic size of some of the RNAi has shown strong size effects (Figure 2.18B) that could be exploited for cell size studies (Cully and Leevers 2006). When conducting a stringent search for genes that are involved in DSB repair, nucleus size can be used

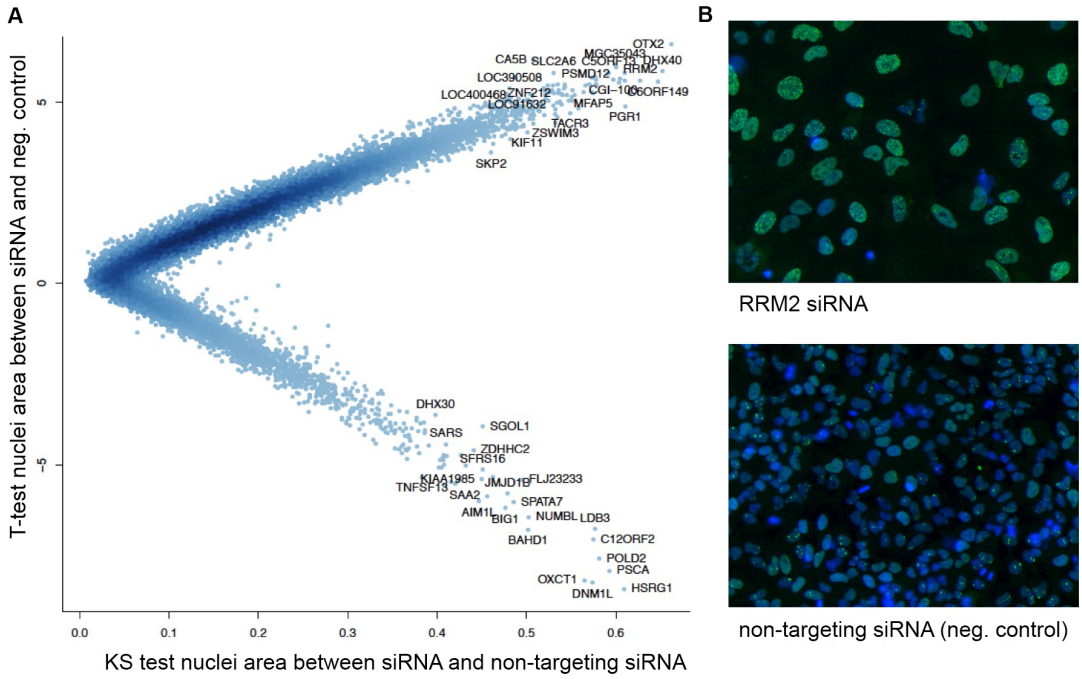


Figure 2.18: Comparison of nucleus size between t-test and KS test based on DAPI. **A** T-test and KS-test scatterplot. **B** DAPI nucleus size difference between RRM2 and non-targeting siRNA.

as a filter to remove unwanted candidate genes. Alternatively, this dataset could be mined for gene silencing that will lead to aneuploidy or loss of heterozygosity.

2.1.7 A Systems view on the DSB screen data

This work has produced a large number of potential genes for follow-up studies. Outliers based on top scoring phenotypes are more likely to play a role in DSB repair. RNF8 and RNF168 were the strongest absolute outliers in our screen and we investigated how these candidate genes are involved in the repair of DSB's (Kolas et al. 2007; Stewart, Panier, et al. 2009). We also characterised a repair complex that had some sequence similarity to a functional repair complex in

Saccharomyces cerevisiae O'Donnell et al. 2010. The complete list of candidate genes with increased number of TP53BP1 foci is given in the supplementary information of ibid. This section provides some high level analysis obtained from all the screening data and related data collected to gain a better understanding of DSBs. I will summarise my findings based on gene enrichment analysis and cross validation with HeLa expression data found in ArrayExpress (Parkinson et al. 2007). As a standard procedure, the gene set was tested for GO enrichment. The annotations for an enrichment analysis were obtained from HPRD. In addition, the dataset was tested using Panther (Mi et al. 2005) and analysed with Ingenuity Pathway Analysis (Qiagen, www.ingenuity.com). Based on the Dharmacon gene identifiers we could map 1,934 genes to perform the GO enrichment. The cut-off was a Z-adjust value > 12 equivalent to 3 times the maximum of the non-targeting negative control 0 ± 4 Z-adjust range. As expected there was enrichment of genes encoding DNA and RiboNucleic Acid (RNA) binding proteins, transcription factors and transcriptional regulators. Surprisingly, we found a large number of genes ($\sim$ 800) that have no annotation but were significant outliers in our screen (see Figure 2.19). An illustrative example of finding protein complexes highly important for DNA repair is the Proteasome (see Figure 2.20) (Krogan et al. 2004). All subunits except for one have shown elevated levels of TP53BP1 foci. An effect on TP53BP1 foci with certain subunits could be reproduced by esiRNA (Figure 2.20). Based on hits that have shown significant changes in TP53BP1 foci number, the Ingenuity pathway analysis found enriched genetic subnetworks shown in Figure 2.21. Several enriched gene cliques were discovered that highlight well established pathways associated with the Ataxia telangiectasia mutated (ATM) signalling network considered, one of the master regulators in DNA damage response (Figure 2.22). In addition, I used ArrayExpress data to validate the expression levels of genes in HeLa cells. My search query contained HeLa and transcription profiling as search terms. After removal of studies with tight research narratives, four studies qualified to be used as sources for HeLa

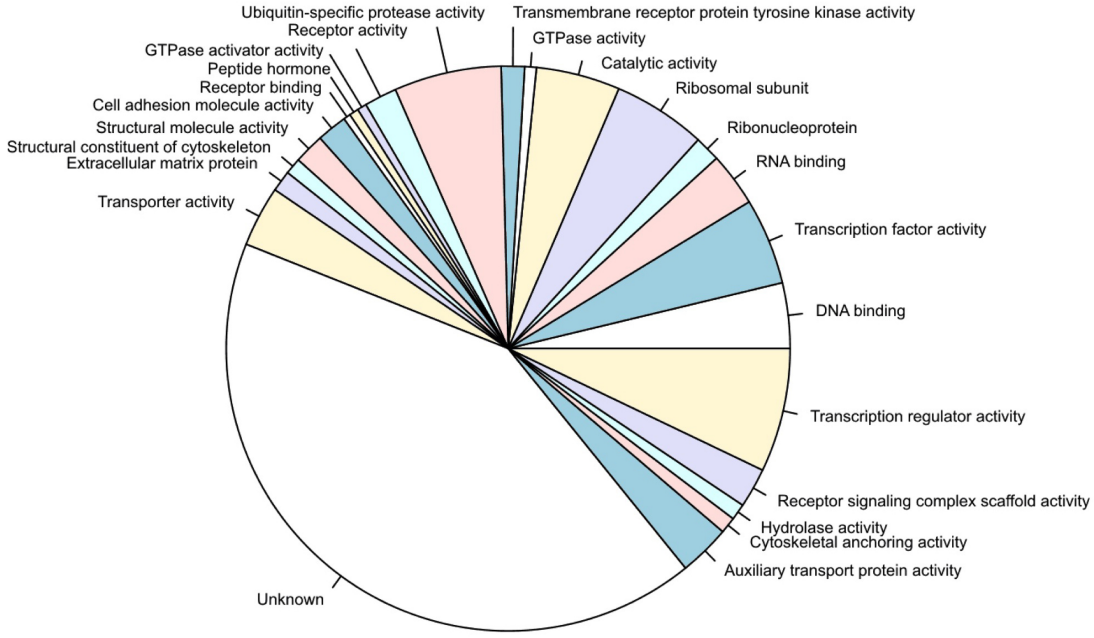


Figure 2.19: Pie chart with enriched GO terms derived from HPRD. Analysis was performed on the 1,934 highest ranked genes in this study that showed phenotypic changes in the HCS. Enriched terms (Bonferroni corrected p-value < 0.01) are shown based on molecular function annotation from HPRD.

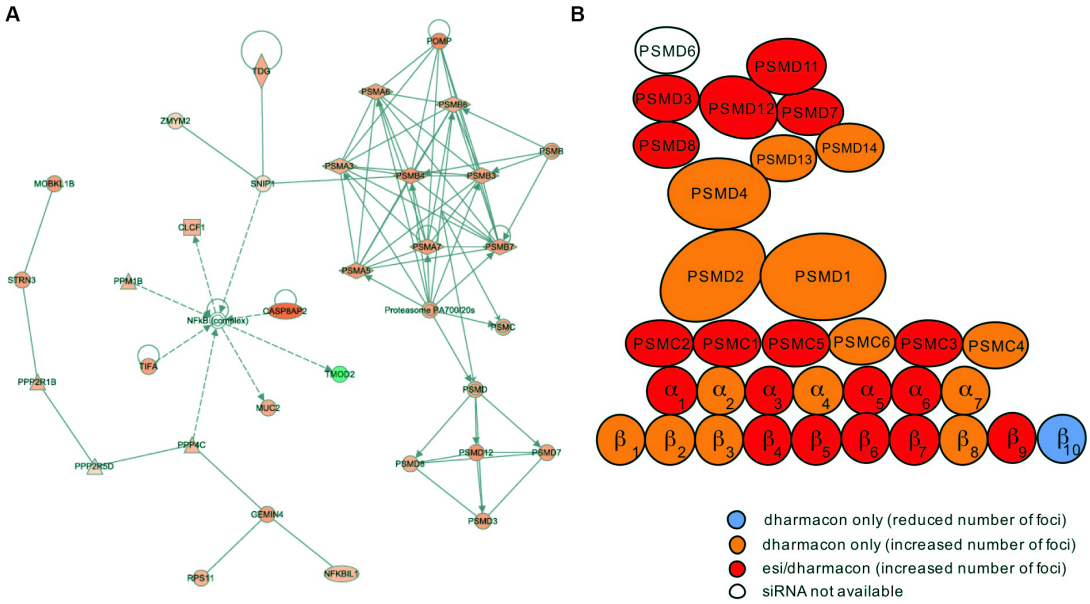


Figure 2.20: Proteasome subunits as mediators of induced DNA DSB's. **A** One of the enriched protein interaction subclusters containing the proteasome subunits. Figure was generated with Ingenuity. **B** All known proteasome subunits are shown. Changes in the number of TP53BP1 foci are represented by colour. Screen was performed with RNAi from Dharmacon or esiRNA.

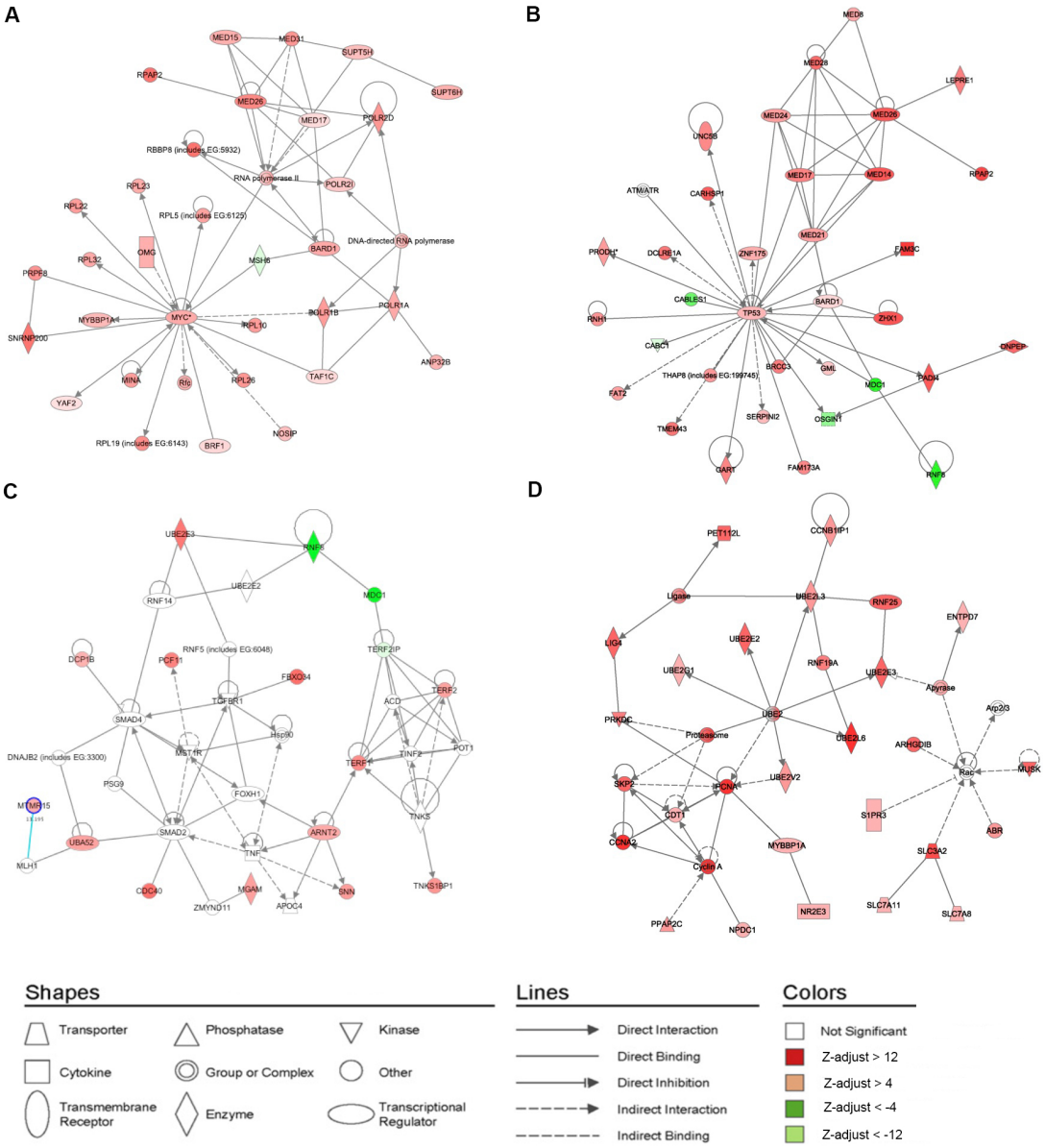


Figure 2.21: Enriched protein interaction networks with increased TP53BP1 foci. **A** MYC protein interaction network. **B** TP53 tumor suppressor protein interaction network. **C** RNF8 related protein network. **D** Core network related to Cyclin A and protein ubiquitination. Network enrichment analysis was performed with Ingenuity.

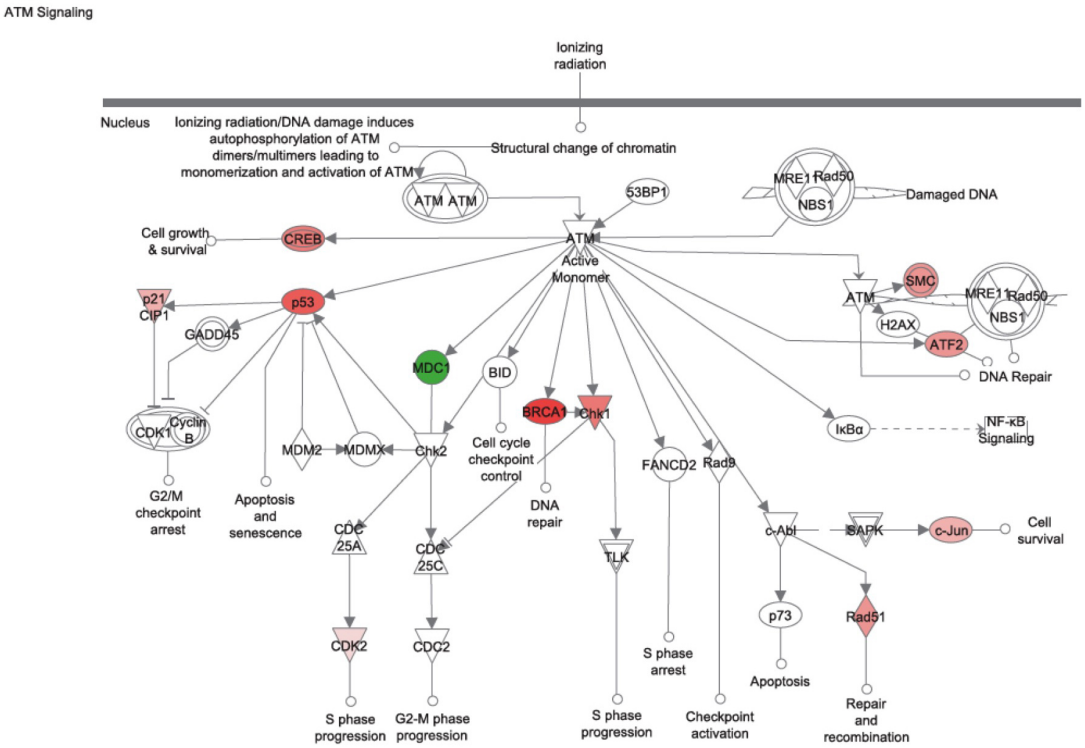


Figure 2.22: ATM Signaling Network. ATM as central element of different signalling pathways. Gene names in red represent siRNA that lead to increased levels of foci, gene names in green represents reduced number of foci. Analysis was performed with Ingenuity pathway analysis.

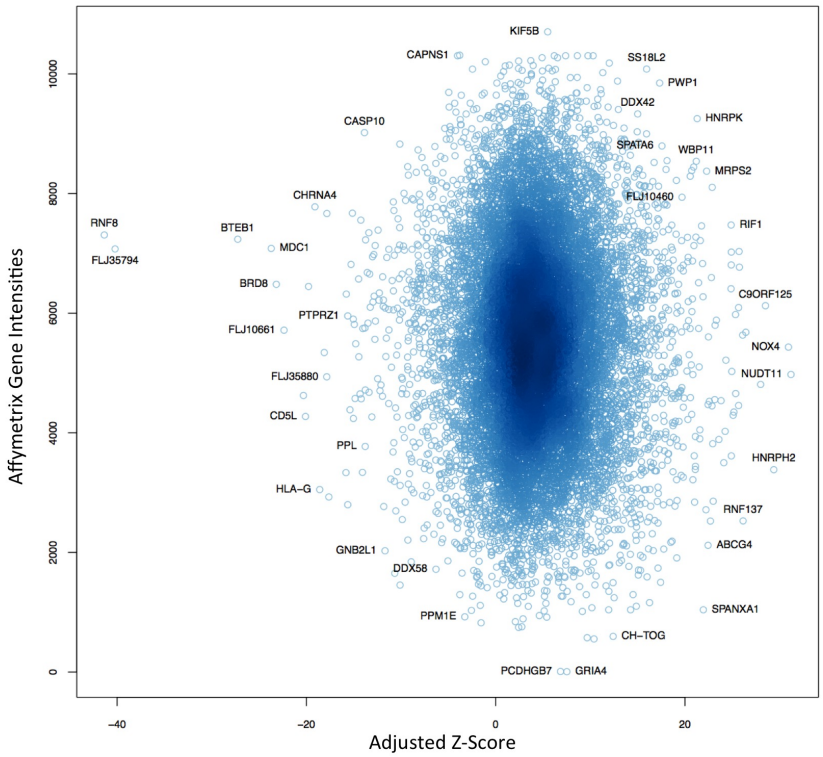


Figure 2.23: Comparing HeLa expression intensities and DSB Z-adjust TSA. Scatterplot between both parameters is indicative of independence between variables.

gene expression data. Today there are 34 experiments available using the same selection criteria. Processed datasets were downloaded and pooled for each gene identifier. The expression differences between the studies were assessed using a t-test to identify outliers and to determine the average expression level and variance for each of the genes. Only 9 genes were excluded based on very low or very high expression out of a total of 12,053 mapped genes. As shown in Figure 2.23, there was no indication of a causal effect between HeLa expression and siRNA activity. Similar results were obtained for expression levels under different conditions like oxygen, hypoxia and targeted silencing of single genes like BReast CAncer 1 tumor suppressor gene (BRCA1).

2.1.8 Multivariate analysis

Before exploring the power of multivariate analysis it is crucial to understand and explore the single parameters that have been derived from an experiment. Multivariate analysis becomes necessary when parameters or variables are conditional and depend on each other. When this project started, it was evident, that we will be able to derive a wide set of parameters with different levels of sensitivity. The foci count initially did not allow us to distinguish between controls. To distinguish the different biological controls, we derived a wide range of parameters to capture the responses in as much detail as possible. In addition to segmentation parameters and their means, medians and quantiles we calculated between-population statistics (for example, between cells with and without TP53BP1 foci for each RNAi). Further, to measure differences between the populations I calculated the KS distance between the diverse parameter distributions as well as the Pearson correlation to quantify the degree of relatedness between pairs of variables. This comprehensive analysis resulted in 70 different parameters. At that stage the visible differences between the density distributions of the parameters looked promising for distinguishing the different controls. Also noticeable were bimodal distributions for some parameters. I used three different machine learning algorithms (a multilayer perceptron, naive Bayes learner and the J48 decision tree algorithm) to assess classification of the accumulated data into the correct sample wells. The motivation to use several algorithms is founded in improved performance using ensembles (Kuncheva and Whitaker 2003). Using all 70 parameters the accuracy for all three machine learning algorithms was above 84% using 5-fold cross validation (see Figure 2.24). To reduce the dimensionality and to emphasise the variation between the variables, I chose PCA and its eigenvectors to display directionality of the variables (Figure 2.9B). Since the controls PCNA, RAD51 and LIG4 clustered in separate groups on each plate, I decided to use the vector length from each median centre of each control pool to every experimental probe

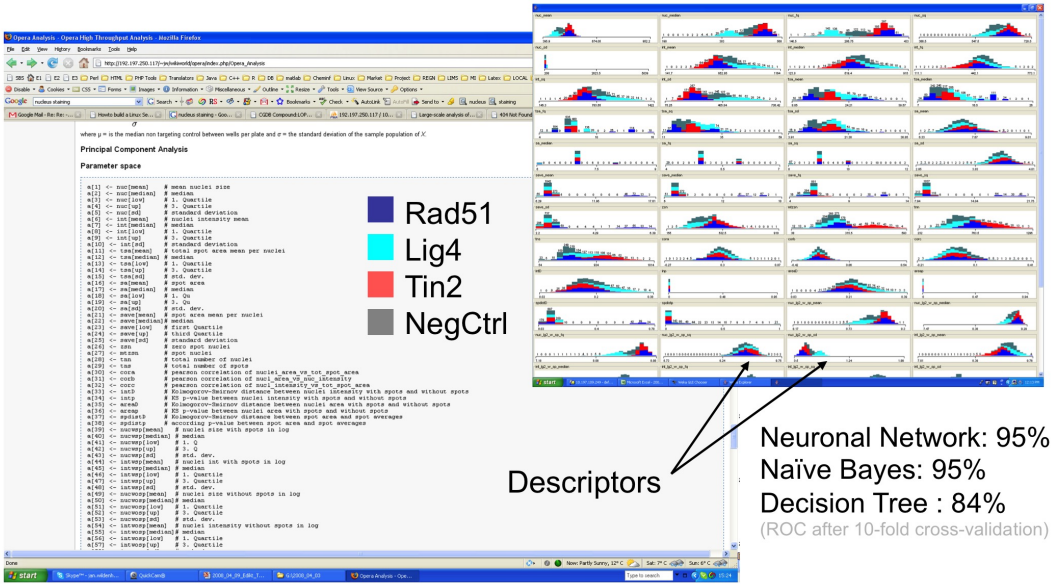


Figure 2.24: Properties derived from image segmentation parameters. Screenshot of project wiki showing property list derived from images. Inset: Display of distributions for some of the properties for four different RNAis: RAD51, LIG4, TIN2 and NCTRL in WEKA.

per plate as a distance measure of phenotypic similarity. The vector length was calculated based on Euclidean distance from the first three components. The hypothesis was that the distance to each control would reflect a phenotypic similarity to each positive control (see Figure 2.25A). Because the orientation of the first components could differ between plates due to extreme outliers the phenotypic similarity measure was calculated on a plate by plate basis. Figure 2.25B highlights the distance for all datapoints to the positive and negative controls. The skewed density kernels follow a β-distribution, with a small distance representing high phenotypic similarity. For example, RAD51 showed moderate levels of increased TP53BP1 signal with little difference to the non-targeting control. However, the phenotypic vector similarity can help to identify functionally linked genes (Figure 2.26A). We were able to identify supporting evidence for genes with functional and physical links to ANAPC10, UBE3C, BUB3, RPL6 and NAT16 (FLJ39237) (Figure 2.26B). Similarly, the left tail of the balanced PCNA control distribution is likely to contain targeted genes with a phenotype that might be functionally related to PCNA. This methodology was similar to (Breinig et al. 2015), where the authors employed a phenotypic similarity measure between single parameters and perturbations with RNAi and double perturbation with two RNAi to explore the up and downstream relationships of silenced gene-gene interactions.

2.1.9 Biological discoveries

The key discovery from the work presented in this chapter was the finding that the ubiquitin ligase RNF8 is essential for the recruitment of TP53BP1 to sites of DNA lesions (Kolas et al. 2007). Initially, the analysis focused on genes that impaired DDR, as indicated by an increased number of TP53BP1 foci in HeLa cells in response to ionizing radiation. The discovery of RNF8 lead to a change

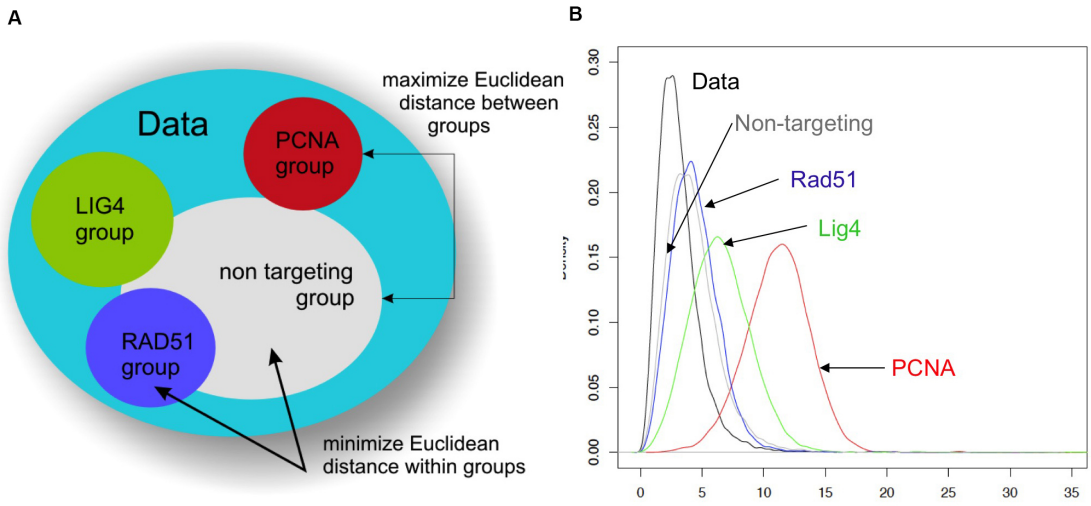


Figure 2.25: **A** Schematic of multidimensional clustering with different control biomarkers. Phenotypic similarity for each siRNA to each control is measured in Euclidean distance. **B** Euclidean distance distributions for each datapoint from the median centre of plate data, non-targeting, RAD51, LIG4 and PCNA cluster.

of the hypothesis because it fully impaired TP53BP1 foci formation, leading to a reduced number of TP53BP1 foci. The assay and data analysis were very robust, so we could confidently classify siRNAs that increased as well as decreased the number of TP53BP1 foci. The RNF8-targeting siRNAs were the strongest inhibitors of TP53BP1 foci formation in the screen with an adjusted Z-Score TSA of -40.2 compared to negative controls. A second paper based on this screen described the role of RNF168, another ubiquitin ligase, in the DDR (Stewart, Panier, et al. 2009). RNF168-targeting siRNAs also significantly reduced the number of TP53BP1 foci (Z-score = -34.4). Prompted by the HCS analysis, experimental studies on RNF168 revealed that RNF8 ubiquitylates H2A-type histones to recruit RNF168 to DSB sites, where it is required for the formation of K63-linked ubiquitin conjugates. Lack of the RNF168-induced chromatin modifications lead to DNA lesions that cause a phenotype similar to RIDDLE syndrome, a recently described human genetic disorder (Stewart, Stankovic, et

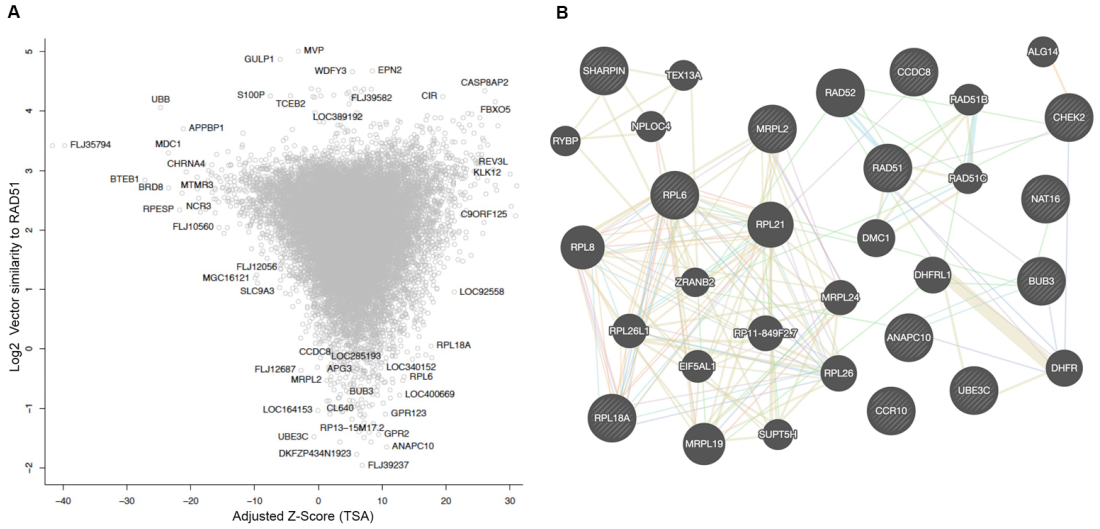


Figure 2.26: Vector similarity and network information between RAD51 and functionally related genes. **A** Vector similarity (Euclidean distance) to RAD51 biomarker compared with adjusted Z-Score TSA. **B** Interaction network containing ANAPC10, UBE3C, BUB3, RPL6 and NAT16 (FLJ39237). Network created from shared protein domains (27%), co-expression data (26%), pathway information (15%), predicted (7%) genetic and physical (5%) interactions using Genemania (Mostafavi et al. 2008)

al. 2007). This work highlights the importance of basic research and screening approaches for understanding human disease. RNF8 and RNF168 cooperate to reorganise chromatin around DSB lesions to recruit TP53BP1 and BRCA1. These two novel ubiquitin ligases thus function downstream of the classical ATM, γ-H2AX and MDC1 effectors of the DDR (Panier and Durocher 2009). The impact of our work on the DDR field is illustrated by the fact that the RNF8 and RNF168 papers have been cited over 900 times in total. Finally, my analysis also initiated the identification of other uncharacterised ORFs that led to an increased number of TP53BP1. One of these ORFs, C6orf167, was confirmed to have weak similarity to MMS22, a known effector of the DDR in budding yeast. MMS22 is required for homologous recombination after stalled DNA replication forks in yeast and our analysis allowed the identification of the gene and characterised the human homolog, called of MMS22L, and its interaction partner TONSL (O'Donnell et al. 2010). The MMS22L-TONSL complex physically interacts with components of the DNA replication fork and is recruited to RPA-bound single-stranded DNA (ssDNA) to promote the loading of RAD51 during Homologous Recombination (HR). Depletion of MMS22L or TONSL results in a marked hypersensitivity to the topoisomerase I poison camptothecin, which is most likely caused by an inability to promote RAD51-mediated repair of broken replication forks. Our results suggest that MMS22L-TONSL is a recombination mediator important for the promotion of genome integrity in S phase.

2.1.10 Comparisons to related work in the DNA repair field

We also compared this DNA repair study with similar DSB repair studies (Paulsen et al. 2009; Doil et al. 2009). First, I compared choices of controls and screen quality (Figure 2.27). The comparison between the studies performed by Paulsen

(Figure 2.27C) and Doil led to Pearson correlation values below 0.1. Studies on the following examples highlighted the importance of the choice of reagents, biomarkers and experimental parameters. For example, the markers γ-H2AX and TP53BP1 are commonly used to study DNA double strand breaks. During pilot screens, both markers were used on two imaging channels. The variability of γ-H2AX compared to TP53BP1 was much higher in our workflow and consequently, γ-H2AX was removed from consideration as an assay readout. A recent study compared both markers in human leukocytes and noted that γ-H2AX signal peaks at 1.5 h post radiation followed by a rapid signal decline. Further, it has since been shown that γ-H2AX exhibits a high level of inter-individual cell variability and was not found suitable for radiation studies (Lassmann et al. 2010). Another genome-wide study of DNA damage used the same HeLa cell line and Dharmacon RNAi library but instead quantified the intensity of γ-H2AX foci (Paulsen et al. 2009). This study yielded a different result set than our screen with a Pearson correlation of 0.02. Most compellingly, the authors of this study identified genes that have been linked to the Charcot-Marie Tooth Syndrome (GJB1, EGR2, PMP22, MPZ, SBF2, MTMR2, HSPB8). However, only GJB1 showed a significant elevation of DNA lesions in our screen. Another similar study on RNF168 used GFP tagged marker proteins (as opposed to antibody detection) in human osteosarcoma cell line (U2OS) cells and siRNAs from Invitrogen (Doil et al. 2009). The results of this whole genome study showed no correlation with our screen (0.03) or the Paulsen et al. screen (0.04). Approximately 50 genes involved in DNA repair (e.g. RNF 168, RNF8, PCNA, RAD51, CHEK1) and genes that encode the alpha and beta subunits of the proteasome were shared hits between the three studies (Figure 2.27C). The silencing of the genes PCNA (Z-score = 26.4), CCNA2 (22.2) and CASP8AP2 (21.5) lead to a severe increase in DNA lesions, whereas RNF8 (-32.1), RNF168, MDC1 (-20.2), UBB (-19) and E2F1 (-18.8) prevented the recruitment of DNA repair proteins to lesion sites. I built a list of uncharacterised ORFs for follow-up investigation in which C6ORF167 (11.8) showed weak homology to

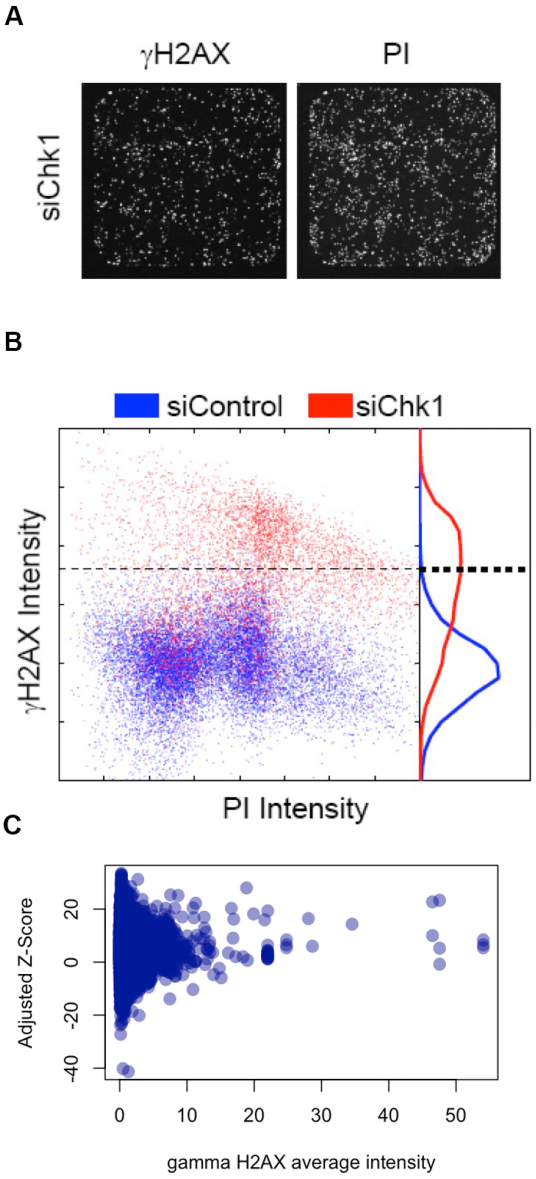


Figure 2.27: Paulsen et al. positive and negative control performance. **A** Images with foci from PI representing cells and γ-H2AX representing DNA damage response. The signal Intensity levels of the foci were used for quantification. **B** Distribution of positive and negative control signals in γ-H2AX and PI intensity. **C** Scatterplot between the average γ-H2AX values and the adjusted Z-Score TSA values.

bakers yeast MMS22. Further experiments helped to characterise the complex of MMS22L (11.8) and TONSL (11.5). This follow-on investigation, based on my analysis, served as a basis for additional follow-up studies. For example CASP8AP2, a strong hit that increased DNA lesions in HeLa cells, has since been investigated in more detail (Hummon et al. 2012). CASP8A2 plays a role in regulating expression of replication-dependent histones and microarray analysis revealed that loss of CASP8A2 function induces the expression of about 2000 genes.

2.2 Application of developed methods to viral infection

I was also involved in an RNAi screen that identified host factors required for VACV replication (Beard et al. 2014). The screen was initially performed on fluorescence intensity plate readers to identify siRNA that either increased or decreased viral load in HeLa cells using the druggable RNAi collection from Dharmacon. Since the first screen did not have imaging data, it was impossible to distinguish siRNAs that affected viral load from siRNAs that affected HeLa cell viability (Figure 2.28A). Therefore, the screen was repeated as an HCS on an Opera microscope. Initially the screen was conducted with a limited number of fields per well. The screen was then repeated to provide a second biological repeat. To assess the quality of the replicates I used the Pearson QC plots as described in 2.1.5. I wrote a customised Acapella script to quantify a range of cellular properties, including intensity, structure and size of objects in three signal channels for nucleus (DAPI), actin structures (phalloidin) and the Green Fluorescent Protein (GFP) expressing vaccinia strain VACV-A5eGFP. The script was optimised to detect and characterise actin stress fibres and actin projections,

as well as primary and secondary virus factories. The actin staining helped to define cell boundaries and detect projections that VACV exploits to infect neighbouring cells (Figure 2.28C). Further I adapted my HCS data management portal for use with the data generated by the VACV screen. A set of 35 parameters was derived from the images (Figure 2.28D). The segmentation procedure to detect actin projections during vaccinia invasion represented a particular image analysis challenge because actin projections extended across cell objects and were highly variable in length and width. The seeded cell density prior to the screen also influenced the formation and shape of actin projections. The broad selection of positive controls provided an opportunity to cross validate the robustness of the segmentation parameters and choices of the biological controls. The HCS used 14 different controls (Figure 2.28E) with putative positive controls Enhanced Green Fluorescent Protein (EGFP), *Herpes virus* DNA polymerase catalytic subunit (ORF28), Sp1 transcription factor (SP1) *herpes virus* transcriptional regulator (ICP4), BCL2 associated athanogene 3 (BAG3), Host Cell Factor C1 (HCF), PRotein Kinase AMP-activated non-catalytic subunit Beta 1 (PRKAB1), Tumor necrosis factor receptor superfamily member 14 (HVEM), Virion protein involved in morphogenesis (VP16), *Vaccinia virus* polymerase siRNA pool 1 (VACVPol1), *Vaccinia virus* polymerase siRNA pool 2 (VACVPol2) and Polo Like Kinase 1 (PLK1) and three negative controls mock, non-targeting siRNA and siRNA which is not processed by the RISC machinery (RSCF). For all control pairs, a Euclidean distance distribution was calculated and the controls EGFP, SP1, ICP4 and ORF28 were chosen to establish a positive cluster median and the negative controls for a negative cluster median (Figure 2.28F). For each plate the distance to all other RNAi on the plate was calculated and a one-sided Wilcoxon rank-sum test was performed. An RNAi was classified as a hit, if the value was in the right tail of the negative control distribution and in the left tail of the positive control distribution (Figure 2.28G). Potential siRNA candidates were identified based on the distance from negative control vector similarity.

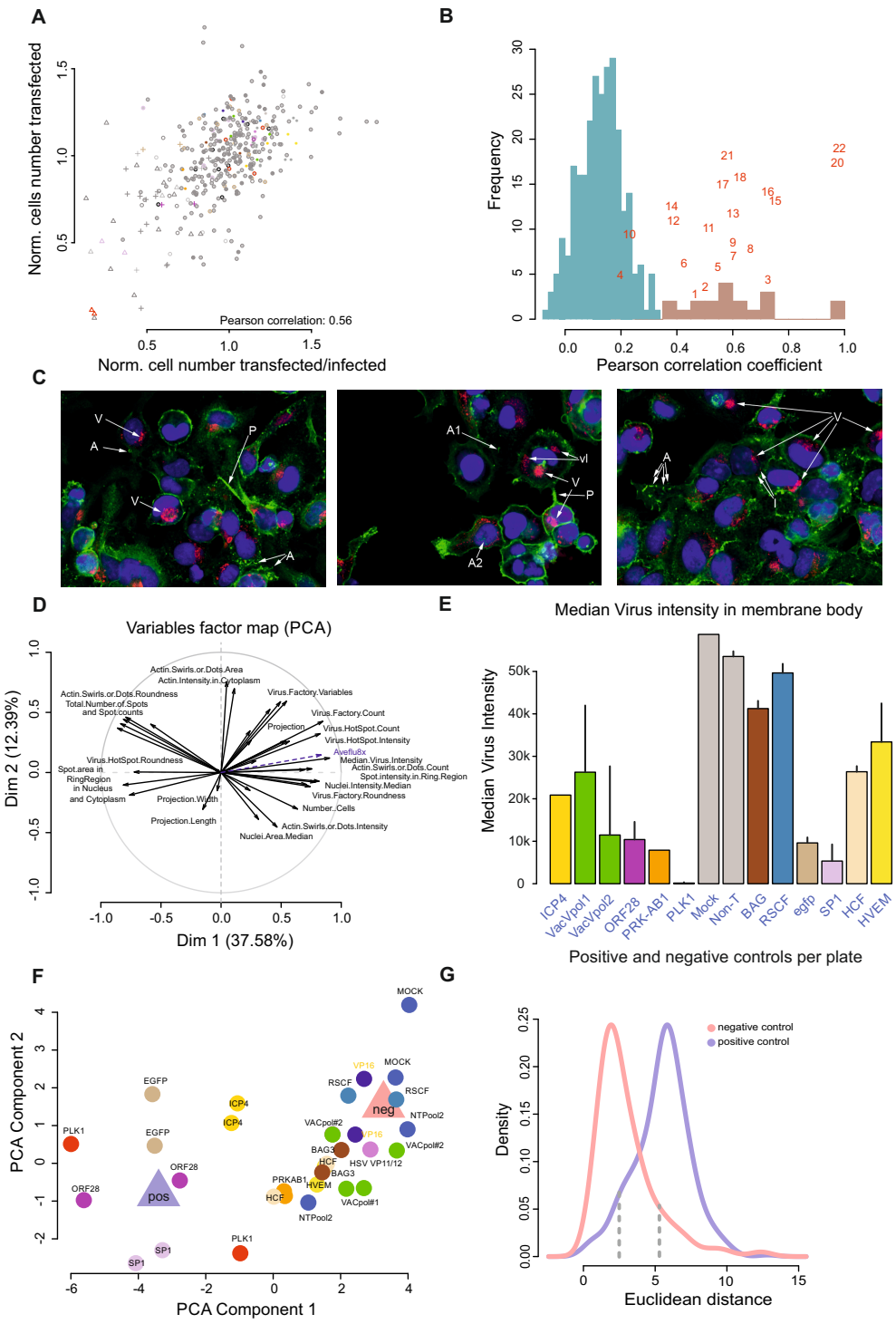


Figure 2.28: VACV HCS summary. **A** Cell count in infected/transfected cells. **B** QC plot for all plate pairs. Numbers represent the plate numbers and the Pearson correlation for the corresponding biological replicate. **C** Opera images of HeLa with VACV infection. **D** Factor map of the 35 parameters. **E** Viral knockdown with controls used in study. **F** PCA shown for positive and negative controls with cluster centres (triangles). **G** Hit selection based on Euclidean distance from cluster centres. Panels A, D, E and F show data from plate 18 between biological replicates.

Additional criteria of the hit selection model were levels of viability similar to negative control in the transfected control screen and reduction in viral load by at least 30%. Changes in actin composition for some siRNA were striking but not an exclusion criteria at the early stage. RNAi for about 800 were considered as outliers, for all candidates images and plate statistics were re-examined and 198 gene selected for experimental follow-up experiments. The list of 198 genes was analysed using STRING, Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis Jr et al. 2003) and Cytoscape (Shannon et al. 2003) for functional analysis (Figure 2.29). Viral load can be reduced through silencing of genes that function on the extracellular, intracellular and translational level. We identified three ADAMTS that belong to the group of metalloproteinases functioning in the ExtraCellular Matrix (ECM). Studies have shown that Matrix MetalloProteinases (MMP)s play a crucial role in the immunoresponse (Elkington, O'Kane, and Friedland 2005). They can be part of the natural immune protection but also known to be hijacked by pathogens to contribute to a pathological state. Like other secreted metalloproteinases, such as several MMPs, it is likely that ADAMTS proteases operate close to the cell surface through interactions with pericellular matrix or cell-surface molecules (Apte and Parks 2015). Further we identified subunits of the proteasome and the ubiquitin system previously shown to play an important role in VACV proliferation (Mercer et al. 2012). Biological processes that affect viral replication include clathrin, Golgi vesicular trafficking translation, transcription, several DNA repair pathways and AMPK complex proteins (Beard et al. 2014). These host cell functions may represent new targets for development of anti-viral therapeutics. I initiated the idea to perform an additional screen to control for changes in actin formation since the HCS analysis revealed that some wells with RNAi and viral load caused severe changes in phenotype. The siRNA library was screened against HeLa cells without viral load to retrieve parameter background estimates for DAPI and phalloidin.

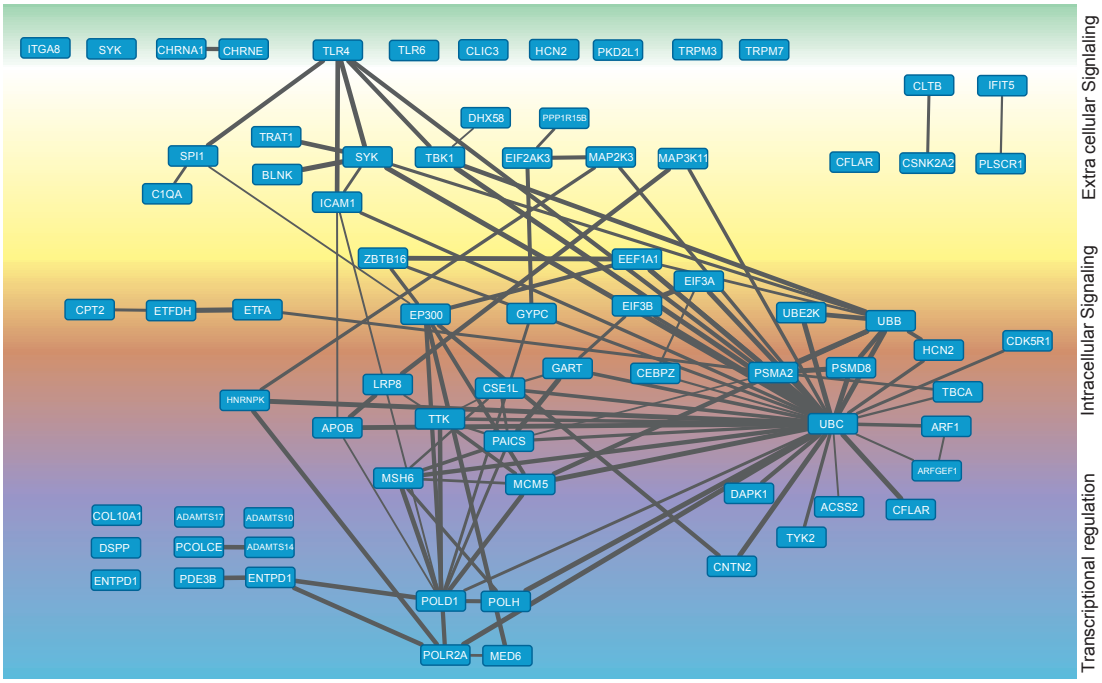


Figure 2.29: Interaction network of RNAi that lead to significant inhibition of viral infection in this study. From top to bottom: Genes known to be involved in extracellular signalling, intracellular signalling and transcriptional regulation. Interaction data was obtained from STRING.

Most noticeable, the cell viability results showed a high Pearson correlation (0.6-0.75) between the infected and non-infected assay plates indicating that the viral load had a minor effect on the viability of cells compared to RNAi exposure. We also observed that if the cell density is low, the infection of neighbouring cells will be reduced. In this context, single cell high content screens with viral loads have been studied in detail by Snijder and colleagues (Mercer et al. 2012; Snijder, Sacher, Rämö, Damm, et al. 2009; Snijder, Sacher, Rämö, Liberali, et al. 2012) who argue that accounting for population context-propagated effects improves reproducibility, cell line comparability, and siRNA phenotype consistency. As an example, these authors cite datasets where the hit list composition changed by up to 50% depending on data analysis procedures. The authors propose a model that considers the fraction of cells that sit on the edge of cell-islets as an exponential function in relation to population size, based on the argument that edge cells and islet cells show different phenotypes and susceptibility to viral infection. In our HCS, viral factory density also increased with cell counts, following a linear relationship (see Figure 2.28B). Hypothetically, the geometric property of the perimeter to the area fosters nonlinearity. However, in our data we did not notice differences in viral load within cell colonies. This discrepancy might be due to differences in magnification, the 10x magnification performed by Snijder, Sacher, Rämö, Damm, et al. 2009 versus the 40x in our study, could provide an observational advantage.

2.3 Recommended statistical procedures for HCS analysis

My research has focused on development of data analysis methods for complex image-based readouts in genetic and chemical high content screens. Each of these

projects used leading edge technologies to address different biological questions. Many issues that I dealt with during assay development in these screens are now commonly part of the analysis workflow in HCS/HTS publications. To provide a simple guide, the following steps are essential and will need iterative optimisation in the initial stages of any screen:

1. Set resolution and image number that yield a robust population size and signal range across several channels to track fluorescent signals and derive cellular features to distinguish biological conditions.

2. Derive as many parameters from imaging data as possible that represent basic segmentation structures (nuclei, spots, intensity levels, orientation), statistical descriptors (mode, median, deviation, kurtosis), algorithmic descriptors (e.g. Voronoi tessellation, ridge and texture characterisation including Gabor and Haralick (Haralick 1979)) and model-based segmentation algorithms (e.g, annotated cell shape segmentations, cell cycle state, projections, cross-channel signal localisation approximations).

3. Check parameters for experimental and temporal bias, spatial effects and dependencies and adjust if necessary.

4. Normalise parameters to address biases with appropriate methods.

5. Assessment of parameter robustness using Pearson correlation, similarity distance metrics or parametric and non-parametric statistical tests (Massey Jr 2012) between each parameter, technical and biological replicates and positive and negative controls.

6. Optimise and group parameter list through factorisation (Feng et al. 2009), principal component analysis (Tanaka et al. 2005), Gaussian mixture models (D. K. Singh et al. 2010; Slacka et al. 2008), linear decomposition (Laufer et al. 2013), clustering (Bakal et al. 2007; Qiu et al. 2011) or machine

learning (e.g. Support Vector Machine (SVM) in (Berger et al. 2008; D. K. Singh et al. 2010) or rule based in (T. R. Jones, Anne E Carpenter, et al. 2009)). Alternatively, create nuisance parameter distributions from different factorisations of the initial variables. The relevant quantities can be treated as estimates of Kullback-Leibler divergence and selected to fit a model (D. K. Singh et al. 2010; Siracusa et al. 2005; Slacka et al. 2008). Another model-fitting approach is to use density dependent sub-sampling by grouping the cells into representative cell sub-populations (Qiu et al. 2011).

7. Secondary statistics for biological hypothesis testing and systems analysis.

2.4 Discussion

An advantage of HCS is the possibility to examine the detailed localisation patterns of proteins in single cells. Cell cultures in HCS are not usually synchronised, which adds a potential rich source of data but at the same time can be a daunting feature to analyse. Since differences in conditions will influence the cell cycle distribution or morphology of the cells in each well, for each experimental design the explicit dependence of a readout on cell cycle state or morphology should be assessed. Cell behaviour can also be affected by cell density in an experimental setting (Snijder, Sacher, Rämö, Liberali, et al. 2012). In stem cells, for example, the expression of cellular markers depends largely on the cell context and exposure to external factors (Ermakov et al. 2012). Proper experimental design and rigorous control of screening conditions are therefore crucial to achieve appropriate seeding densities. In addition, monitoring single cell attributes and correlating data for cells in close proximity to each other can yield valuable insights. However the complexity of HCS data requires refinement cycles of parameters that can be derived by detection algorithms and the statistical analysis applied to them.

Technological advances allow ever higher throughput screens and the monitoring of multiple signals will extend multi-parametric phenotyping. These increasingly complex datasets will need detailed statistical analysis to identify meaningful experimental results. For example, recent studies from the Boutrous group have begun to use HCS of RNAi treated cells to derive epistatic relationships (Laufer et al. 2013; Fischer et al. 2015). This is a huge undertaking since it requires not only to compare an experimental condition (knockdown of two genes by RNAi) against standard controls, but also against individually perturbed states (knockdown of single genes by RNAi). These multiple comparisons reduce the statistical power of the HCS data. The authors addressed this by adapting the experimental design by focussing on a small number of genes representing key pathways. A similar approach to study endocytic activity and associated organelles by high content imaging allowed identification of regulatory interactions and hierarchical structures (Horn et al. 2011; Liberali, Snijder, and Pelkmans 2014). Approaches that combine single cell measurements, data for multiple parameters and genetic interactions have been reviewed recently (Liberali, Snijder, and Pelkmans 2015). A second major issue in the field is the availability of central repositories for high content data, their data structures and analysis. To take full advantage of imaging technologies will require development of common standards and tools to store and perform comparative analysis for different datasets. In other areas of biomedical research there are established resources that are maintained by the National Institute of Health (NIH) (PubChem[3]) and European Molecular Biology Laboratory (EMBL) (ChEMBL[4]). In addition, resources such as the Connectivity-Map (Lamb 2006) have been built by the Broad Institute to explore functional connections between gene expression datasets. In chemical biology and drug discovery, sources such as PubChem and ChEMBL serve this purpose. Resources for imaging data are however more problematic. HCS imaging data is

[3]https://pubchem.ncbi.nlm.nih.gov/
[4]https://www.ebi.ac.uk/chembl/

storage intensive and the infrastructure needed to run algorithms on the stored information are computationally intensive. An interesting parallel is the Google image data repository, for which classification algorithms compare images and find similar images with annotated content. An application of this technology to problems in HCS should therefore be feasible. A specific argument for the need for such a resource can be drawn from viral infection studies (Snijder, Sacher, Rämö, Liberali, et al. 2012) and stem cell research. Many different studies use the markers Octamer-binding transcription factor 4 (OCT4) and Nanog Homeobox transcription regulator (NANOG) to assess stem cell state and it would be beneficial to use existing data to extend and support new data and findings. Yet it is currently not feasible to compare similar studies at the level of primary data analysis. In my opinion, it would seem reasonable to set up a collaborative resource to develop standards for image acquisition, to collect and centralise imaging data, and to build an object segmentation and cell classification algorithm repository. There are several algorithm repositories that provide feature extraction for commonly present morphological assays (Held et al. 2010; Anne E Carpenter et al. 2006; T. R. Jones, Kang, et al. 2008; Hu and Murphy 2004; Schneider, Rasband, and Eliceiri 2012). Data format standards for HCS that capture images with processing parameters such as HDF5 data format for cell-based assays (CellH5) (Sommer, Held, et al. 2013) and SDCUBE (Millard et al. 2011) are actively developed, while building a phenotype catalogue of cell morphologies to categorize cell structures has been suggested (Abbas, Dijkstra, and Heskes 2014). Automated morphological profiling has been showcased for localisation detection (Berger et al. 2008; Collinet et al. 2010; Koh et al. 2015) and cell shapes (Yin et al. 2013). A recent study reported the PhenoPlots visualisation tool for analysis of morphological features derived from HCS data (Sailem, Sero, and Bakal 2015). This tool enables researchers to relate quantitative values to images through pictorial representations of up to 21 individual features. The power of information acquisition from HCS data could

be leveraged when algorithms for data dimensionality reduction and Machine learning (ML) could be seamless lyintegrated. Another idea would be automatic recognition of phenotypes with immediate reporting of occurrences with cell structures previously not registered. In summary, given the increasing complexity and prevalence of HCS data sets, standards for statistical analysis procedures, phenotype ontology for cell morphologies and data storage solutions are essential.

Chapter 3

ML supported cheminformatics to leverage chemical HTS

The aim of this chapter is to present approaches ranging from applied statistics, machine learning to cheminformatics that I have taken to analyse different data sets to discern novel bioactive compounds. The discovery of novel bioactive molecules is a challenge because the fraction of active candidate molecules is usually small and confounded by noise in experimental readouts. Cheminformatics can improve robustness of chemical high-throughput screens and functional genomics data sets by taking structure-activity relationships into account. We successfully applied Quantitative Structure Activity Relationship (QSAR) to find that phenothiazines and apomorphines can act as regulators of cell differentiation in murine embryonic stem cells (Diamandis et al. 2007). Further, we pioneered computational methods for the identification of structural features that influence the degradation and retention of compounds in the nematode *C. elegans* (Burns, Kwok, et al. 2006). I also used cheminformatics to assemble a comprehensive screening library of Previously Approved Drugs (PAD) for redeployment in new bioassays (Ejim et al. 2011). The combination of chemical genetic interactions,

cheminformatics and Machine learning (ML) enabled us to predict novel synergistic anti-fungal small molecule combinations from chemical-genetic interactions (Spitzer, Griffiths, et al. 2011; Wildenhain, Spitzer, Dolma, et al. 2015). In a study on biological effects of commonly prescribed psychoactive compounds, we were able to show a strong link between lipophilicity and bioactivity of psychoactive compounds in yeast and their functional genetic responses that could account for unwanted side effects in humans (Ericson et al. 2008). We extended this investigation to reveal Structure Activity Relationship (SAR)s in chemically diverse compound collections that were used to probe chemical-genetic interactions in yeast (Hillenmeyer et al. 2008). Some of the methods and tools are available to the scientific community, including an open source software package called MolClass allowing researchers to make predictions about bioactivity of small molecules based on their chemical structures (Ishizaki et al. 2010; Castonguay et al. 2015; Wong et al. 2013; Wildenhain, Fitzgerald, and Tyers 2012).

3.1 Assembly of chemical libraries

There are different strategies to exploit small molecule activities to probe biology. Chemical libraries can consist of well-established molecules with known drug targets or completely new chemical matter. Molecules that are well characterised can provide phenotypic markers or new insights with the potential to be repurposed for new indications (Paolini et al. 2006; A. L. Hopkins 2009; Keiser, Setola, et al. 2009; Huang et al. 2011). The PAD library was developed in collaboration with a leading Canadian screening centre at McMaster University by identifying drugs already present in the in-house collections and purchasing additional compounds from chemical vendors. Further, to explore new chemical matter we identified a collection of yeast bioactive compounds within a large 53,000-member

synthetic library of previously uncharacterised molecules. From this data I assembled a dedicated Yeast Bioactive Compound Library (YBCL), which has found many applications in different species and target-based assays (Wildenhain et al., manuscript in preparation).

3.1.1 Creation of a compound collection to repurpose approved drugs

The Nobel laureate James Black famously stated: The most fruitful basis for [discovering] a new drug is to start with an old drug (Raju 2000). Repurposing of drugs became a new paradigm, in part because the magic bullet approach, originally conceived of by Paul Ehrlich has only produced a relatively small number of approved drugs (Strebhardt and Ullrich 2008). Work with approved drugs reduces the barriers to regulatory approval and significantly shortens the time it takes to bring drugs to the market. The idea of new uses for old drugs has flourished in the academic community (Chong and Sullivan 2007). The systematic combination of drugs in the clinic, whether arrived at by ad hoc tests or rational design, has yielded various successes in different areas of medicine, including cancer, malaria, HIV and neurological disorders. The company CombinatoRx, for example, focused on the systematic discovery of two-component therapeutics based on reference-listed drugs (Borisy et al. 2003; Lehár, Zimmermann, et al. 2007; Lehár, A. S. Krueger, et al. 2009).

To systematically perform combination testing, we assembled a small molecule library of 1270 PAD. This was achieved by systematically curating existing drug lists from governmental organisations such as the U.S. Food and Drug Administration (FDA) and the World Health Organisation (WHO). For this purpose I developed a web-crawler collecting information form websites to build compound

databases and filters using Python and chemistry Application Programming Interfaces (API's) including Open Babel (O'Boyle et al. 2011) and Frowns[1]. The assembly of the PAD library is outlined in Figure 3.1. Non-redundant PADs were cherry-picked from a sub-collection of known bioactives in our in-house collection, sourced from 1,120 Prestwick compounds (Prestwick Chemical Inc, Illkirch, France), 1,267 Sigma Library of Pharmacologically Active Compounds (LOPAC) compounds (Sigma-Aldrich, Oakville, Ontario), 2,000 Microsource Spectrum compounds (Gaylordsville, Connecticut), 499 natural products from Biomol (Plymouth Meeting, Pennsylvania) and 63,249 Maybridge compounds (Cornwall, UK). We matched these compounds in two ways, based on structure and via their generic name. The structures were obtained from Medbase (unpublished, 2,400 compounds) and Drugbank[2]. All molecules were stripped of salts and checked for uniqueness. The list of generic names was compiled from the current WHO[3] and FDA[4] drug lists.

This library has been used for a drug screen that discovered enserazide and loperamide as synergistic enhancers of the antibiotic minocycline (Ejim et al. 2011). It was also used to find synergistic combinations with flucanozole (Spitzer, Griffiths, et al. 2011).

[1] http://frowns.sourceforge.net/
[2] http://www.drugbank.ca/, 1,177 compounds
[3] http://www.whocc.no/atc_ddd_index
[4] http://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs

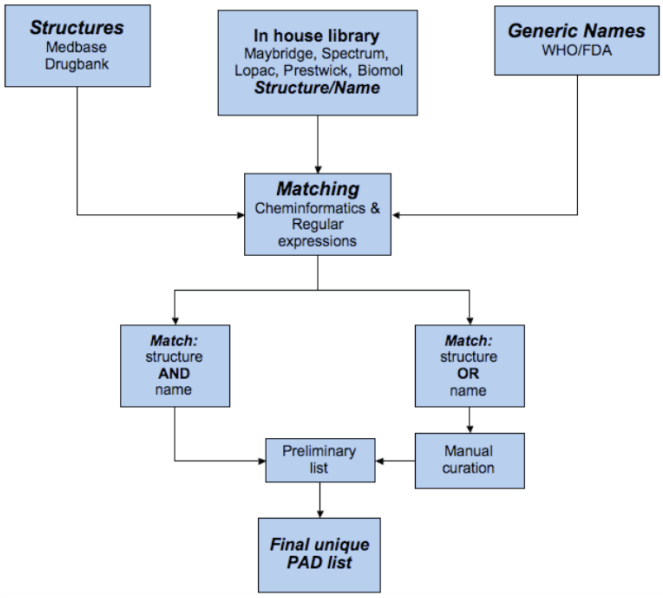


Figure 3.1: Workflow of PAD library assembly. Molecules with available structure and drug names were matched against the curated in-house databases and compound libraries. If the search had a correct match, plate and well information were collected and an approved drug record was created in the PAD. If only the generic name of a drug structure was matched, molecules were curated against several sources such as PubChem (Y. Wang, Bolton, et al. 2010), ChEBI (Degtyarenko et al. 2008) and DrugBank (Wishart 2008) and the data was added to the in-house database. If the selected molecules and structures were approved drugs and available in the in-house compound repositories, the information was added to the PAD.

3.1.2 Exploration of new chemical matter with a Yeast Bioactive Compound Library

Expanding the molecule drug space has become a challenge over the last decades. Currently, there are about 1,552 unique approved drugs on the market (Law et al. 2014). James Black's paradigm still dominates todays medical chemistry and research usually relies on historical confidence in existing scaffolds to design new molecules. Similarly drug companies became careful with renewed enterprises into forward chemical screening. The war on cancer in the 70's encouraged the chemistry divisions in pharmaceutical companies to synthesise millions of molecules and to test them in reporter assays focused on known cancer pathways and tumor proteins. The output of these endeavours did not hold up to the expectations and efforts slowed down significantly by the end of the century. With the 21st century academic research transitioned towards greater automatisation and small molecule libraries became affordable for many research labs. Combinatorial libraries and high throughput screening are highly regarded in academia since they provide a steady stream of new tool compounds, powerful modulators of biological function and starting points for medicinal chemistry. A broad screening strategy has been suggested to explore a large defined compound space (Gordon et al. 2002) and initiated diverse forward screening efforts by many labs to find small molecule candidates for phenotypic discovery. Further, chemical probes, compared with genetic approaches can function rapidly, are reversible and titratable (Kitano 2007). The wide range of available model organisms allows systematic exploration of changes in phenotype and application of chemical probes at different stages of development. The idea to join efforts in academia and use a small set of screening facilities for early hit identification was believed to encourage cross-disciplinary communication within multiple scientific and technological disciplines (Oprea et al. 2009).

The broad screening strategy is one of the core objectives of the National Institute of Health (NIH) (ibid.) and the Joint European Compound Library initiative (Besnard, P. S. Jones, et al. 2015). Within the financial premises of an academic lab, we have chosen a commercially available screening collection that has been shown to be a highly diverse library at the time of purchase (Krier, Bret, and Rognan 2006). Together with a postdoctoral fellow in our group, I performed the analysis of a chemical screen with a 53,000 synthetic molecule library[5] against a drug pump-defective budding yeast strain (Figure 3.2A). The data was processed with methods described in Chapter 2 to analyse High Throughput Screen (HTS) data. LOcally WEighted Scatterplot Smoothing (LOWESS) regression was used to remove a row-column bias evident across the whole screen using an empirically estimated sliding window of 1/3 of the data. Median normalisation was used to normalize between plates, technical replicates ($n = 2$) and biological replicates ($n = 2$). Outliers between technical replicates were excluded from the final analysis as well as outliers between biological replicates, if neither was considered active using a median and Interquartile range (IQR) by a fitted normal distribution with $N(1, IQR)$. The cutoff of 4 Median Absolute Deviation (MAD) was chosen to define bioactive outliers. For all 2,858 inhibitory molecules we generated path fingerprints (length 7) using Python and the Frowns library. Tanimoto coefficients were calculated between all compounds to assess compound similarities. By applying a nearest neighbour search, the most active compounds within a cluster with at least 0.95 similarity were kept, reducing the final library size to 1,570 molecules.

To highlight differences in compound diversity and coverage of chemical space, multidimensional mapping using MultiDimensional Scaling (MDS) or Principal Component Analysis (PCA) are frequently applied to condense a high-dimensional

[5]Maybridge Screening Collection http://www.maybridge.com/

descriptor space into a representation that is accessible to human interpretation (Haggarty 2005; Hillenmeyer et al. 2008; Kutchukian et al. 2016). To observe how the library compares to approved drugs available by the FDA we computed a range of molecule compound properties with MolClass (Wildenhain, Fitzgerald, and Tyers 2012) for the YBCL and DrugBank (Wishart 2008) (Figure 3.2B). The analysis was performed with ten physical chemical properties including molecular weight (MW), number of rings (NR), hydrogen bond acceptors (HA), hydrogen bond donors (HD), rotatable bonds (RB), AlogP and the number of functional atoms that are important in drug structures such as nitrogen (NC), oxygen (OC) and sulphur (SC).

This library has subsequently been screened in many different model organisms including: *Danio rerio* (Ishizaki et al. 2010), *Saccharomyces pombe* (Castonguay et al. 2015), *Saccharomyces cerevisiae* (Wong et al. 2013), *Ostreococcus tauri*, *Arabidopsis thaliana*, *Caenorabdities elegans*, HeLa cells and the pathogens *Candida albicans*, *Cryptococcus neoformans* and *Staphylococcus aureus* (Wildenhain et al., manuscript in preparation). Even though the bioactive compound set was selected based on growth inhibition in budding yeast, this collection produced a wide range of phenotypes in different organisms with up to 10-fold increased hit rates, emphasising the utility of small molecules as probes of conserved biological pathways (Figure 3.3).

Notable are the lower hit rates in *E.coli* and *C. elegans* compared to the other species. For gram-negative bacteria it is difficult to overcome the additional barrier protecting the organism from the environment. Further, *E. coli* have numerous drug pumps and metabolic enzymes that can actively remove or metabolise

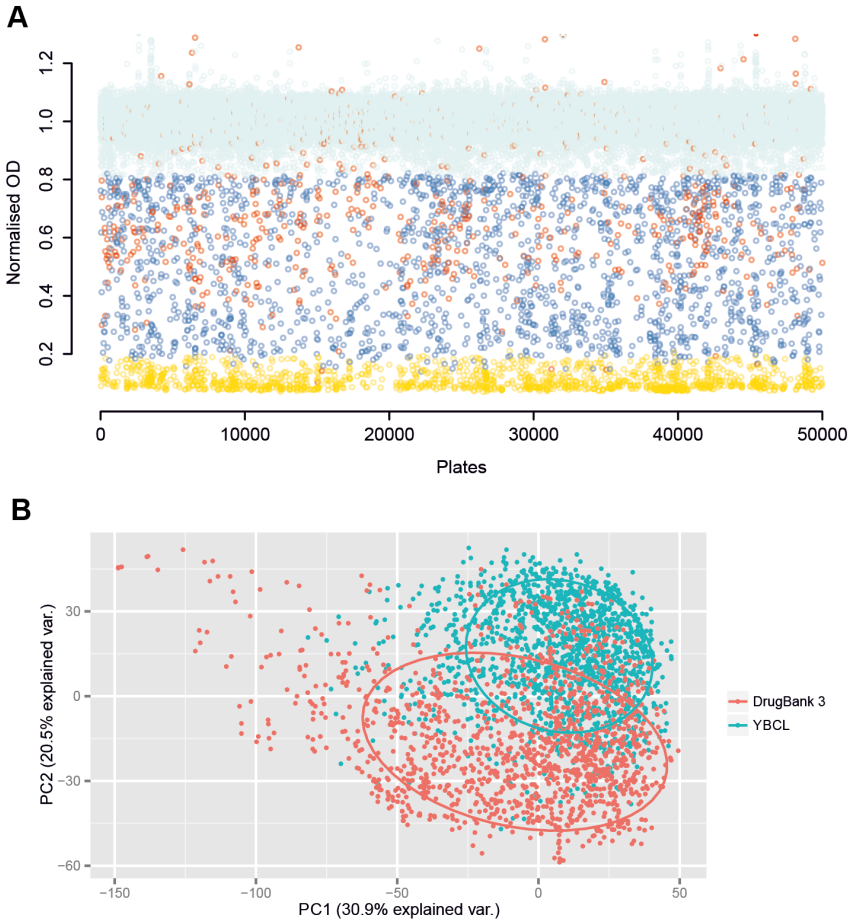


Figure 3.2: Maybridge collection as starting point for the YBCL. **A** Growth of a drug pump-deficient yeast strain in the presence of the Maybridge Hit-finder collection of 53,000 molecules after normalisation. Colours represent non-active (grey), bioactive (blue), highly active (yellow) and non-replicating (red) molecules. **B** Visualisation of chemical space for FDA approved drugs and YBCL using the first two principal components. Properties used for the PCA analysis are MW, AlogP, HA, HD, NR, OC, NC, SC and RB.

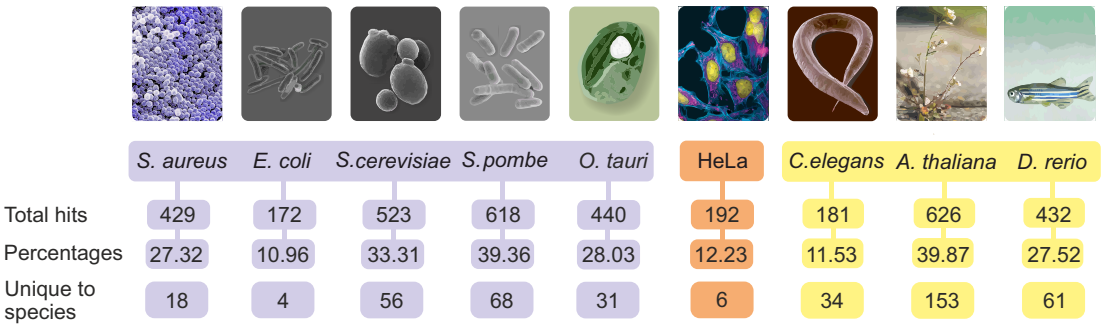


Figure 3.3: Cross Kingdom activity analysis in YBCL. Colours indicate screens performed as quantitative growth assays (purple), *in vitro* assay (orange) and phenotypic *in vivo* screens (yellow). Shown are the number of total hits, the corresponding percentage of active compounds and the unique number of compounds active in one species or cell line.

small molecules. The low hit rate in *C. elegans* can be attributed to the experimental setup since compounds are dissolved in media with *E. coli* that serves as a food source for the worms.

In follow-up experiments, we identified molecules with unique activities using less than half of the initial screening concentration (20 μM) for all species. Previous studies have suggested *in vivo* models to help preselect molecules for drug development. Computational methods are highly cost effective compared to experiments but also bear the risk of missing interesting candidates (see Section 1.1.2 for details). Known methods highlight chemical space limitations such as Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET), receptor binding G-Protein-Coupled Receptor (GPCR) (C. Lipinski and A. Hopkins 2004) and bioaviability (C. A. Lipinski et al. 2001). Property filters for tool, lead and drug like compounds depend on expectations (Workman and I. Collins 2010; Wallace et al. 2011) which will lead to different enrichment rates. To assess how different molecules performed in the different assays we compared molecule properties that seem important for activity in the different species (Figure 3.4). The data suggests that the mean of the chemical properties changes in each of

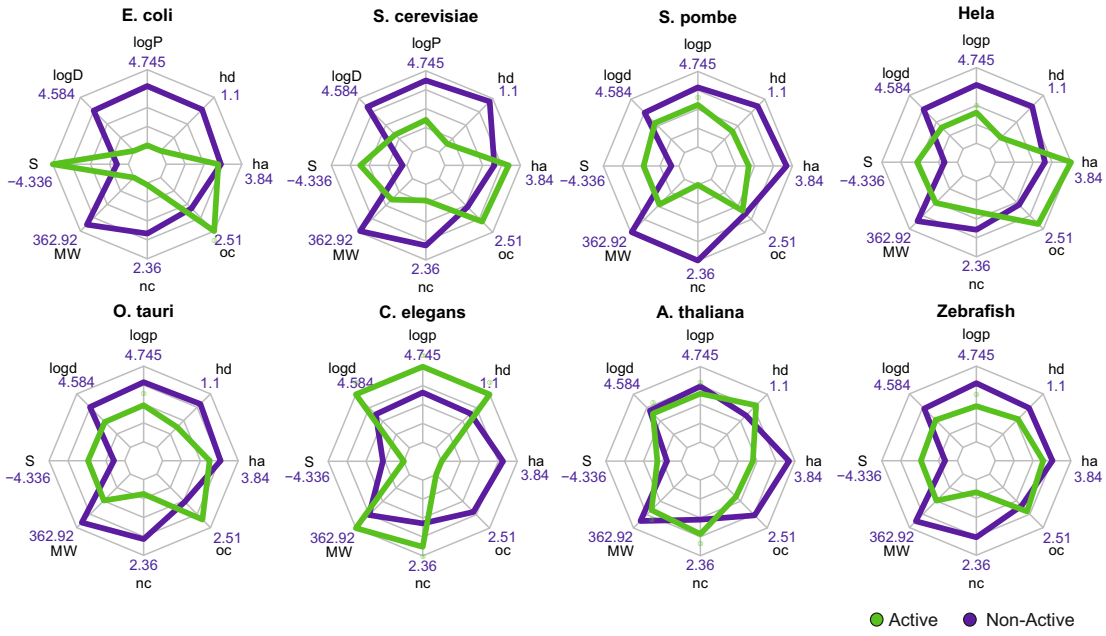


Figure 3.4: Chemical properties responsible for small molecule activity shown as radar plots. Compound property means for active (green) and non-active (blue) small molecule fractions in the YBCL. Property labels in clockwise order are LogP, HD, HA, OC, NC, MW, Solubility (S) and LogD. Grey lines represent 60% to 100% of the values shown for each property.

the species. Unsurprisingly, *E. coli* shows increased sensitivity to small soluble molecules and lipophilic molecules have been shown to be effective sensitisers in gram-negative bacteria (Ejim et al. 2011; Zabawa et al. 2016). The opposite seems to be true for *C. elegans*. The complexity of targeted small molecule screens in worm will be discussed in more detail in Section 3.2.3. Further, our results suggest that molecules that contain nitrogen are more likely to show activity in *A. thaliana*, whereas Henrietta Lacks cervical cancer (HeLa) cells are more susceptible to molecules with hydrogen bond acceptors and oxygen.

The absolute mean difference between the organisms are small. Some evidence for this is provided in the consistency of the non-active molecule background

distribution for the different species. Still, we can find molecules that are active in a single species and many that are active across several species (at the concentration screened). Drug discovery literature focuses on criteria such as permeability, potency, and selectivity for potential lead like molecules. Promiscuous effects have been a burden for the pharmaceutical industry but have also proved valuable information to find novel therapeutic applications. The low number of selectively active molecules indicates the difficulty to find targeted, non-promiscuous compounds.

Additional screens highlight the power of our YBCL library. A screen in fission yeast revealed two molecules that disrupt heterochromatin-mediated gene silencing through inhibition of the Clr3-containing Snf2/HDAC repressor complex (Castonguay et al. 2015). Alleviation of gene silencing was also observed in *Arabidopsis thaliana* and a mouse cell line. A high-throughput assay in budding yeast identified four molecules that inhibit human telomerase (Wong et al. 2013). We also characterised novel chemical features (Chenoweth 1956) that lead to a copper deficiency phenotype in zebrafish and were not previously annotated with metal chelation (Ishizaki et al. 2010). The same screen also identified several 5-nitrofuran compounds that affect melanocyte development in zebrafish embryos. Mode of action studies revealed aldehyde dehydrogenase to activate 5-nitrofuran compounds in several species (Ishizaki et al. 2010; L. Zhou et al. 2012). These findings have important implications for managing toxic side-effects of the prodrug nifurtimox which is used to treat bacterial and trypanosome infections. The compounds in our synthetic library were built using combinatorial chemistry and could be re-synthesised with modification to the core scaffolds to explore QSAR. Such information was used to perform a library analysis to obtain candidate molecules from commercial vendors and through synthesis (Wong et al. 2013; Castonguay et al. 2015). We believe the YBCL library contributes to a broadened understanding of compound specificity and promiscuity of previously

uncharacterised molecules and will lead to improved ADMET models for public and academic community resources.The next two subsections will present two further screens with the YBCL: analysis of time-series data and High Content Screen (HCS) data generated with the YBCL.

Small molecules as modifiers of circadian rhythm

In collaboration with Laura Dixon from the Millar laboratory at the University of Edinburgh the YBCL library was screened in *Ostreococcus tauri* to measure changes in the circadian rhythm. Circadian clock biology has been increasingly important in a wide range of organisms ranging from cyanobacteria to mammals. To build models for clock mechanisms and to correlate function, the shape of the oscillation of a postranslational clock protein can be measured. Many key components in the clocks of different species have been discovered by forward genetic approaches. Reverse approaches using known drugs that inhibit different pathways have shown altered circadian properties (O′Neill et al. 2011). Clock profiles can provide insight into the mode of action of molecules as inhibitors of enzymes that can be conserved across different kingdoms. *O. tauri* is a good clock model since the light-dark cycle can be measured by CCA1-LUC expression. In the preparation of the screen, cells were entrained for 6 days in light-dark cycles. On day six, media was refreshed and luciferin added. On day 7, 2 μL of compound library were added with a liquid handling robot (2 μM final concentration) and plates were moved to data collection. Data was recorded over one light-dark cycle and then four days of constant light (see Figure 3.5A).

Data was collected in batches of ten plates, each plate containing four different controls: DMSO, cordycepin (CORDY), cycloheximide (CHX) and 3-(3,4-dichlorophenyl)-1,1-dimethylurea (DCMU), in quadruplicate wells. These controls represent the vehicle response DMSO, period lengthening with CORDY,

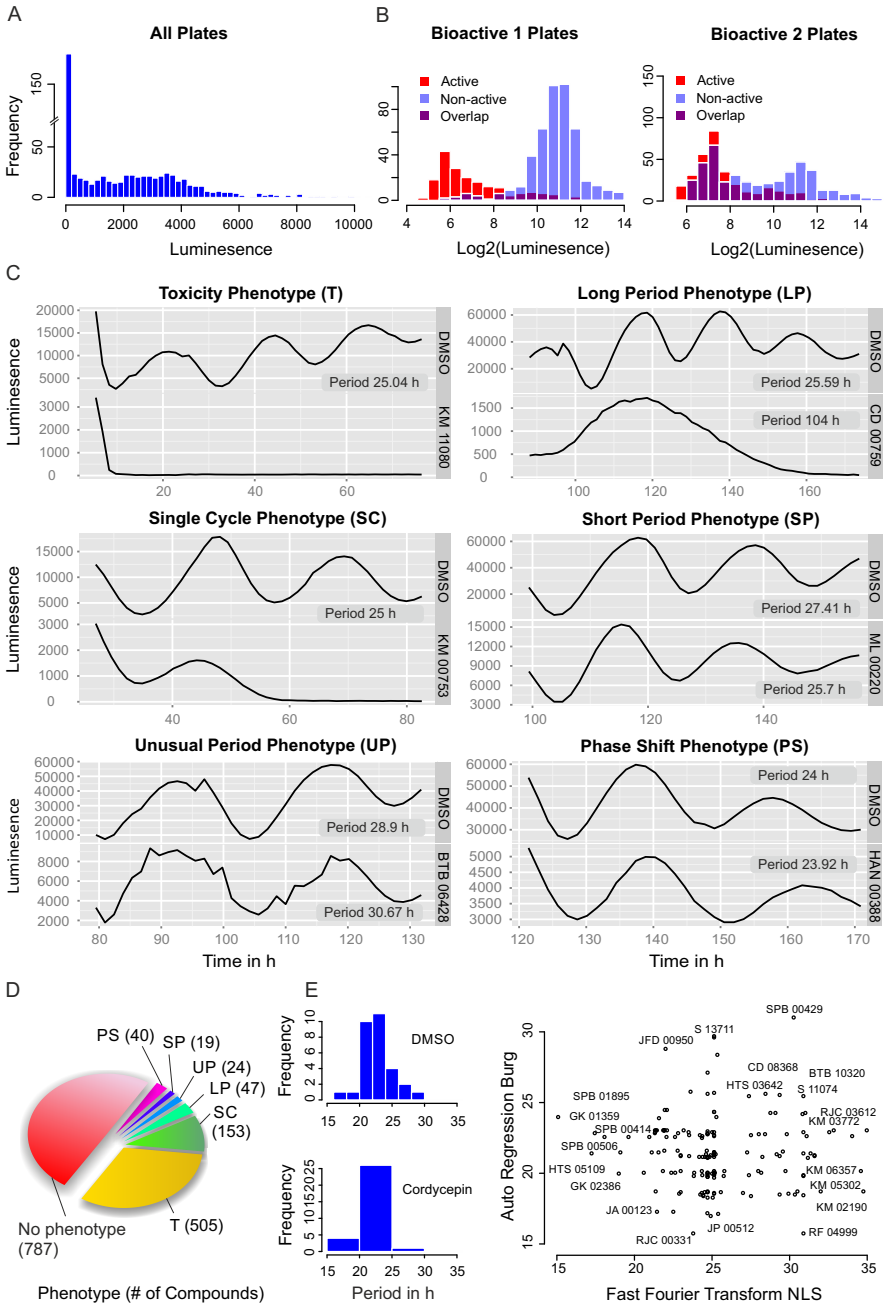


Figure 3.5: Viability and circadian rhythm summary from *O. tauri* data. **A** luminescence of CCA1::LUC cells 15 hours after initiation. **B** Distribution of the Log2 luminescence signal of active and non-active compounds in bioactive and cytotoxic YBCL plates. **C** Examples of circadian periodic signals for compounds that cause toxicity (T), long period (LP), single cycle (SC), short period (SP), unusual period (UP) and phase shift (PS). **D** Pie chart with proportions of phenotypes. **E** Distribution of period length for cordycepin and DMSO controls. Scatterplot comparing estimated period lengths using the Burg and FFT-NLS algorithms.

period shortening with CHX and loss of viability DCMU. The screen was conducted with two biological replicates. The viability of cells was determined based on luminescence signal (see Figure 3.5B). Compounds were classified as toxic (T) if they caused a response similar to DCMU: the median luminescence was below 250 cps between 15 - 20 hours into experimental initiation and the luminescence has only decreased from the time of compound addition. Further, I characterized compounds that interrupt luminescence within the first cycle (SC) as late responders and compounds that caused unusual periods (UP) as bioactives (see Figure 3.5C). Molecules that showed a prolonged period (LP), a shortened period (SP) or a phase shift in period (PS) compared to DMSO and CORDY controls were also classified as bioactive (see Figure 3.5D). The period of the CCA1-LUC expressing cells was analysed over a free run with continuous light between 36 and 120 hours. To estimate the circadian period length I fitted an AutoRegressive (AR) time series using the Burg method (Burg 1967), Chi-square periodogram (Sokolove and Bushell 1978) and FFT-NLS (Straume, Frasier-Cadoret, and Johnson 2002) to get three estimates for the period (see Figure 3.5E). I developed an R web application using shiny (Timeseries.R) for the analysis of time series data and it is available at: http://sysbiolab.bio.ed.ac.uk:3838/my-app/TimeSeries.shiny/.

A high-content imaging DNA damage screen with the YBCL

The response to DeoxyriboNucleic Acid (DNA) damage has been considered a target for cancer therapy since the early days in the 'war on cancer' (Nixon). In response to DNA damage, cells activate cell cycle checkpoints and the DNA damage repair machinery to secure genomic integrity. If the DNA damage is severe and cannot be repaired by the cell, affected cells undergo apoptotic cell death. Since the rate of cell division in cancer cells is higher than in normal cells, cancer therapeutic agents target predominantly the DNA Damage Repair (DDR) pathways. Further, tumor cells are often deficient in one or more DNA repair

pathways and tend to be dependent on the remaining intact repair pathways for survival (Farmer et al. 2005). We have shown Tumor Protein p53 Binding Protein 1 (TP53BP1) to be a powerful marker to monitor repair of Double Strand Break (DSB) since they represent the most severe form of DNA damage (Kolas et al. 2007). We screened the YBCL using the HCS approach outlined in Chapter 2.1. The screen outcome is summarised in Figure 3.6.

None of the derived parameters resembles known distributions (Figure 3.6A-E), emphasising the need for non-parametric tests (Chapter 2.1.6). The experimental parameters showed a weak correlation between the TSA and TSI values and the G_1 and G_2 assignment of nuclei (Figure 3.6G). We summarised the properties to four relevant markers of cell count for viability, nucleus size, DNA repair and cell cycle arrest (Figure 3.6H). Cell viability in response to small molecules does not need to be linked to DNA damage (Figure 3.6I). Only one third of the molecules used in this study reduce cell numbers and affect the DSB response marker. Figure 3.6J summarises the quality of the screen. There is a good correlation between the two replicates and separation of the negative (DMSO) and the positive control (etoposide at 3 concentrations 0.1 μM, 1 μM and 10 μM). We identified 116 molecules that increase the TSA of TP53BP1 foci. About 300 molecules inhibit the DNA damage repair response, possibly due to interference with early signal transduction events or mechanisms that act on processes downstream of DDR. For example, SPB 06987 ($C_{15}H_{14}N_4OS_2$) a hit unique to HeLa cells, has also been found in a DNA re-replication screen that supports DNA repair initiation (see PubChem BioAssay IDentifier (AID) 624296 for details).

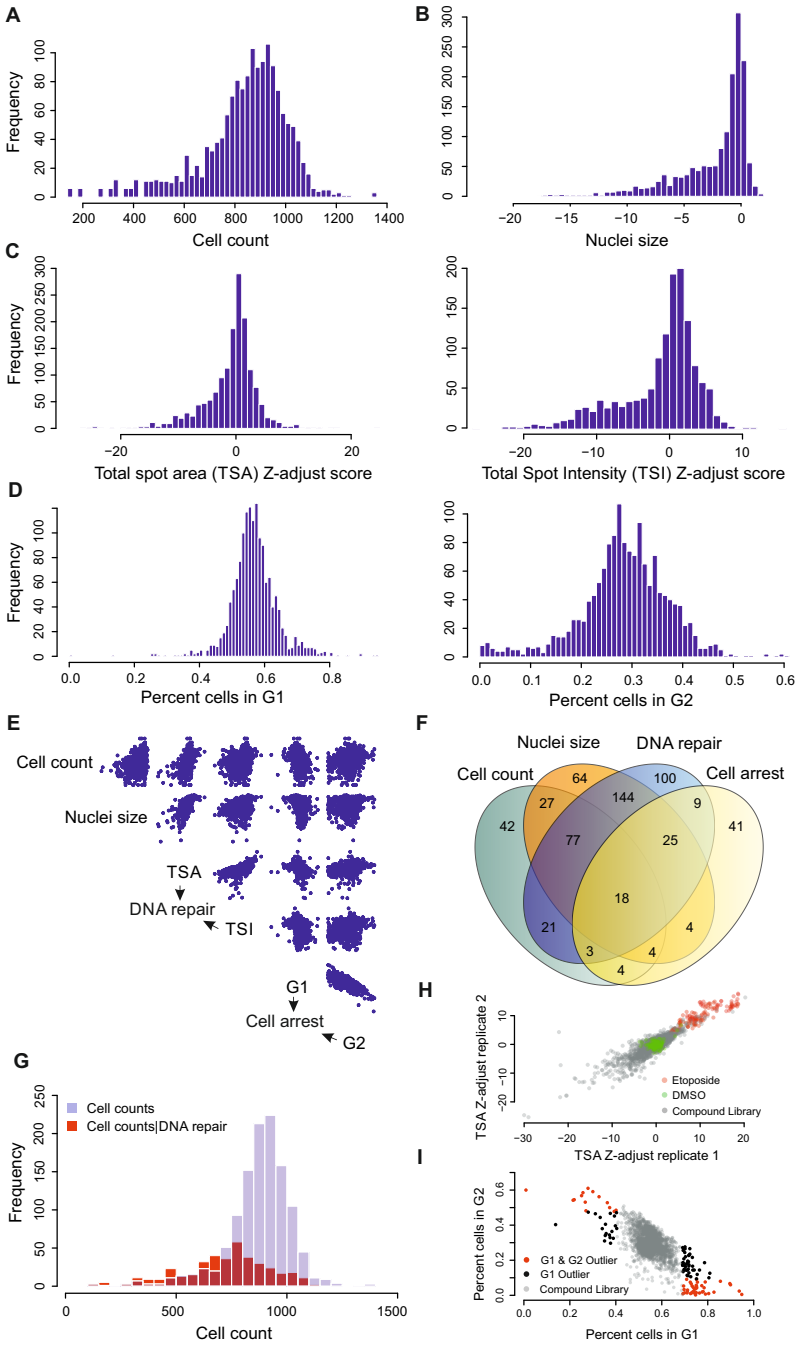


Figure 3.6: Summary of high-content imaging DNA damage screen. **A** Distribution of cell counts. **B** Distribution of nucleic size compared to DMSO control. **C** Distribution of adjusted Z-Scores for TSA and TSI. **D** Distribution of cells in G_1 and G_2. **E** Scatterplots between the six parameters. **F** Venn diagram showing the overlap between phenotypes viability, nucleus size, DNA damage response and cell cycle arrest. **G** Distribution of cell counts (viability) for cells with DNA damage. **H** Replicate plot for adjusted TSA Z-Score values. **I** Scatterplot between percentage of cells in G_1 and G_2 phase.

3.2 Identification of structural features that mediate specific activities

A primary challenge in high throughput screening is to identify promising chemical scaffolds. Structure-activity relationships are important for choosing candidate molecules for follow-up studies. In a typical screen, most scaffolds are not active, some are promiscuous, and ideally one or more will be selective towards the screening target or pathway. Cheminformatic algorithms for example, allow the implementation of automatic filter algorithms to detect pan assay interference compounds (Baell and Holloway 2010). To improve the success rate and the overall knowledge-base, relevant data has to be collected and clustered to segregate interesting molecules and find relevant structural features (Campillos et al. 2008; Keiser, B. L. Roth, et al. 2007; Wildenhain, Fitzgerald, and Tyers 2012). I have demonstrated the application of such methods to identify compounds that affect neuronal stem cell fate and to characterize properties of small molecules that bioaccumulate in *C. elegans* (Burns, Kwok, et al. 2006; Diamandis et al. 2007). The folowing sections will illustrate procedures I developed for targeted molecule classification to identify novel lead compounds with structural features that avoid the recognition of drug pumps in *E. coli*, modulate cell fate decisions of neural stem cells and accumulate in *C. elegans*.

3.2.1 Characterisation of multi drug resistance pumps with ML

Even after much progress and technology-driven development in the field of medical sciences, infectious diseases remain the second leading cause of mortality worldwide. The emergence of Multi-Drug Resistant (MDR) strains of pathogens

led to the resurgence of infectious diseases. Strategies to overcome the problem of drug resistance include the discovery of new drug targets and finding novel compounds that show no cross-resistance to the available antibiotics. Multi-drug efflux pumps are a main cause of drug resistance and pose a serious problem in the treatment of various infections (E. D. Brown and Wright 2016). AcrB pumps found in gram negative bacteria, including *E.coli*, provide a powerful mechanism of defence against a variety of antibiotics. The AcrB protein is a member of the Resistance-Nodulation-Division (RND) superfamily found in all domains of life. The AcrAB efflux system genes of *E. coli* form an operon and are composed of an RND-type transporter and a periplasmic accessory protein, AcrA. It spans both the inner and outer cell membranes and is responsible for extruding a variety of antibiotics, leading to multi-drug resistance by over-expression of the efflux pump proteins. The AcrAB system is unusually broad in its substrate specificity and expels cationic, neutral and anionic substrates (Nikaido and Zgurskaya 2001; E. W. Yu et al. 2003; Hobbs et al. 2012). Studies conducted on *E. coli's* multi-drug resistance pumps apply to similar pumps found in other organisms, like *S. cerevisiae*, that could lead to the development of better treatments for a variety of fungal infections. *S. cerevisiae* is protected against antibiotics by a transport mechanism that is part of the Pleiotropic Drug Resistance (PDR) network (Ernst et al. 2010). Cancer cells can also be resistant to multiple classes of chemotherapeutic agents when proteins are upregulated that belong to the superfamily of ATP Binding Cassette (ABC) transporters (Teodori et al. 2002). To overcome MDRs, much work has focused on developing more potent and selective modulators that could overcome or reverse resistance.

The initial project developed from a collaboration with the Brown lab at McMaster University in Canada and was the first attempt to use ML and compound properties as multivariates to classify and predict small molecule

action. As a starting point, the Maybridge Hitfinder collection of 8,640 molecules was screened against MC1061, a hyper-permeable and streptomycin-resistant strain of *E. coli* (X. Li et al. 2004). Two strains of MC1061, one with an empty overexpression vector and one with overexpressed AcrB were screened against a total of 1,920 hit compounds that were identified in this primary screen. The assays were carried out in duplicate using eleven different compound concentrations ranging from 250 μM to 0.24 μM. Negative and positive controls were included on each 96-well plate, DMSO and a mixture of ampicillin (100 μg/mL) and chloramphenicol (25 μg/mL). After preparation of the plates, the OD_{600} of each plate was measured. Plates were then incubated at 37°C, at 85% humidity for ∼20 hours, after which, OD_{600} was measured. Percent residual growth $\%G$ was calculated and transformed to a semi-logarithmic scale to calculate the EC_{50} from an estimated Hill equation:

$$\%\widehat{G} = \frac{E_{max}}{1 + \left(\frac{I}{EC_{50}}\right)^{c}}$$

where E_{max} is the extrapolated $\%\widehat{G}$ in the absence of an inhibitor, I are the concentration measures in μM of the tested compound and c is the Hill slope coefficient (Gesztelyi et al. 2012). After removing all compounds with incomplete or low quality EC_{50} determinations from the dataset, a total of 1,475 molecules were used to generate and validate the classification model (Figure 3.7). Fold suppression for the 1,450 compounds was calculated as the ratio of the average EC_{50} value in the *E. coli* MC1061-AcrB strain and that in *E. coli* MC1061 (Figure 3.8). As classification algorithm we chose to employ a Naive Bayes (NB) learner. For each feature, it calculates the occurrence probabilities per category from a training set (e.g. compounds pumped or not pumped by AcrB). The most significant of these features classify samples from a test set into their respective category (Xia et al. 2004). The most accurate prediction for the

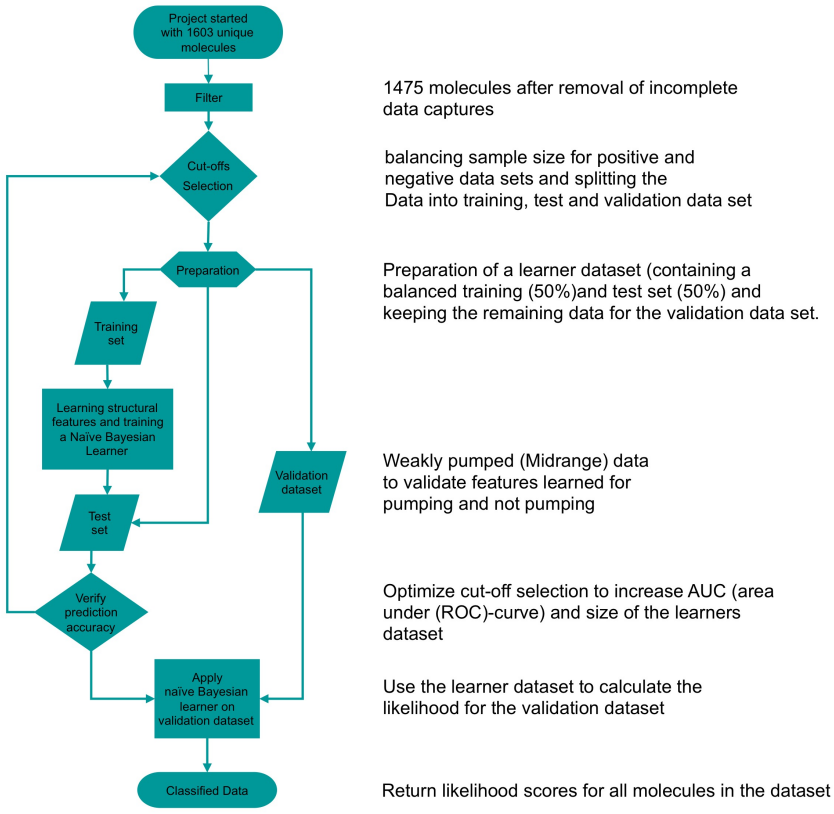


Figure 3.7: Overview of the analysis of the AcrB dataset. Filtering steps removed fits below $R^2 < 0.9$ for replicate molecules with an $EC_{50} > 250\mu M$ and incomplete measurements. Cutoffs were optimised after several iterations. A final dataset with 554 molecules, 266 pumped and 288 not pumped by AcrB, was chosen as a learner dataset. The model was tested using a 50/50 training and test set. The model was evaluated on the remaining 921 compounds representing weakly pumped compounds to verify the likelihood of outliers indicative of molecules to be pumped or not pumped within this set to relax the previous cutoffs.

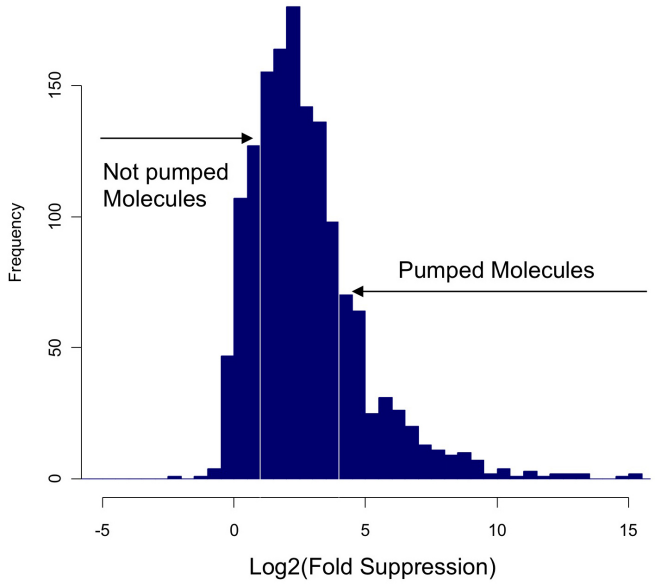


Figure 3.8: AcrB fold suppression histogram of pumped and non-pumped molecules. The cutoffs for non-pumped and pumped molecules are shown. The fold suppression range between these cutoffs is used as validation data.

model was obtained when the fold suppression cut-offs were set to 20 and 2 for molecules that are pumped (266 compounds) and not pumped by AcrB (288 compounds). The data was randomly split into a 50% training set to train the NB classifier and a 50% test set to assess the accuracy of the prediction model. The final selection of 554 molecules made up the learner dataset. In addition to molecules with quantitative fold suppression values, we added 130 molecules that showed complete inhibition in the pump deficient strain. Based on the generated AcrB substrate model, likelihood scores were calculated for the remaining 921 compounds (weakly pumped by AcrB) as a second validation dataset to assess strong likelihood scores in both categories.

The molecular descriptors included AlogP, MW, the number of HD and HA, the number of RB, the Polar Surface Area (PSA) and functional class fingerprints Functional-Class Fingerprints (FCFP) with a path length of up to six atoms. These are structural fingerprints that represent each compound as a series of

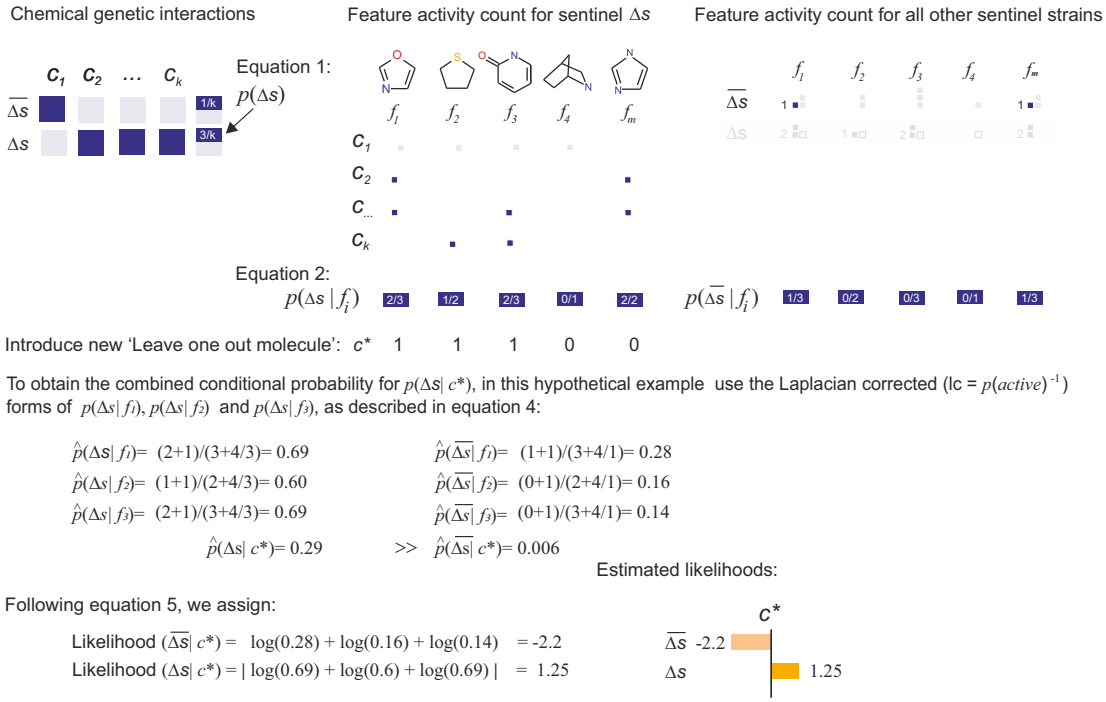


Figure 3.9: Example of naive Bayes learning in cheminformatics. The illustration shoes the two classes in the AcrB model (pumped/not-pumped). The feature count (probabilities) is estimated across all chemical features present in the training in either of the two class data sets. Molecules that are pumped are represented by class Δs and molecules that are not pumped by class $\overline{\Delta s}$.

fragments starting from each heavy atom and extending out up to six bond lengths. FCFP's take into account the function of the atom within the molecule, whether the atom is aromatic, a halogen, a hydrogen-bond donor or acceptor, positively or negatively ionisable, regardless of the atom type (Hassan et al. 2006). Using the Naive Bayes classifier, features that are over- and under-represented in the compounds received positive and negative scores based on calculated probabilities. The overall score for each molecule is calculated by taking the sum of all of the fragment scores (Figure 3.9). Structural features were learned for both classes using the Pipeline Pilot implementation of a Laplacian-Modified Naive Bayes learner. For each class Δs a subset of a compounds will

show activity out of a total of k, this defines the baseline probability of compounds being pumped:

$$p(\Delta s) = \frac{a}{k} \tag{3.1}$$

Each compound contains a number of structural features that are either pumped or not pumped. The conditional probability for each feature f_i that is present in compounds to be active in Δs, would then be:

$$p(f_i \mid \Delta s) = \frac{a_{f_i}}{k_{f_i}} \tag{3.2}$$

The 'naivety' of this approach treats all features as independent probabilities, so Bayes' rule can be written as:

$$p(s \mid f) = \frac{1}{\prod_{i=1}^{m} p(f_i)} p(s) \prod_{i=1}^{m} p(f_i \mid s) \tag{3.3}$$

where $p(s) = a/k$ and m is the total number of features in the training data. A correction parameter is introduced to avoid overconfidence of $p(s \mid f) = 1$, a subset of unique features are all found active in the dataset, and to avoid $p(s \mid f) = 0$, a single feature f_i may not be present. The probability estimate can be approximated and the formula becomes:

$$\hat{p}(s \mid f) = frac(F_s + L * p(s))(F + L) \tag{3.4}$$

where $L = p(s)^{-1}$ is the Laplace correction and F_s is the active feature count in

Δs and F is the total feature count for f_i. The likelihood score E_s is the class outcome given a compound:

$$E_s = \sum_{i=1}^{m} log\Big(\frac{F_s + 1}{F + L}\Big) \tag{3.5}$$

where i sums over all logarithmic feature probabilities 1 to m. The validation for class Δs pumped or not pumped, given a compound c^*, can be performed using leave-one-out cross-validation as illustrated in Figure 3.9.

The Receiver Operating Characteristic (ROC) curve for the AcrB substrate model demonstrates a high proportion of True Positives (TP) and True Negatives (TN) with an Area Under the Curve (AUC) of 0.89 (Figure 3.10). The likelihood distributions for each class suggest that prediction for non-pumped molecules are more prone to False Positive (FP) and therefore it is a challenging task to assign the structural features that prevent them from being pumped by AcrB. The molecules pumped by AcrB show a strong single modality implying a defined set of structural features that contribute to their pumping phenotype. To better understand what the underlying structural features are that lead to the preference of AcrB, we examined physical properties of molecules using PCA (Figure 3.11). The Learner dataset (n=554) was tagged as either pumped or not pumped to provide qualitative class assignments. The fold suppression value was used as a quantitative class measurement to arrange the data points along the x-axis. Approximately 60% of the variation in the data is contained in the first two principal components which are sufficient to drive the separation between molecules that are either pumped or not pumped by AcrB (Figure 3.11A). To highlight which molecular properties are influencing the separation, Figure 3.11B depicts the PCA eigenvectors. The AcrB Fold suppression value is influenced by the number of Halogens, ALogP and the molecular solubility. Less influential in terms of explained variability are HA,

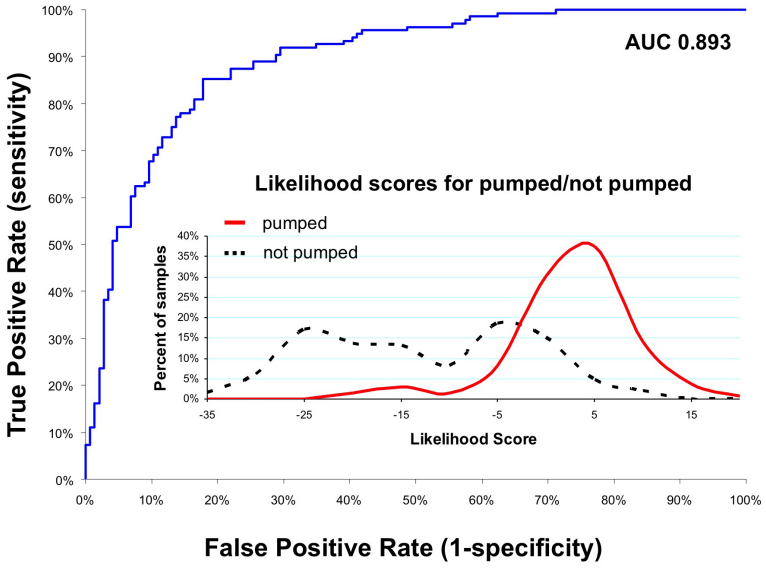


Figure 3.10: ROC curve for AcrB substrate model. Inset: Likelihood distribution of the pumped (red) and non-pumped molecules (black dotted) in the test set.

MW and PSA. Table 3.1 summarises the descriptor median and the ranges that will influence the likelihood of whether or not the molecules are pumped by AcrB. The data suggests that small molecules that are highly hydrophilic are less likely to be pumped than large hydrophobic structures. Further, halogens seem to be recognised and are actively removed from the organism. Molecular descriptors allowed to separate between both classes, posing the question if certain scaffolds are more susceptible to AcrB transport.

We calculated the Bemis-Murcko fragments (Bemis and Murcko 1996) for the molecules (Figure 3.12). Similarly we examined the NB feature probabilities of substructures that are frequent in one class and not the other to understand how this affects the molecule transport. The data suggests that structures with more than two joined rings are more likely to be pumped.

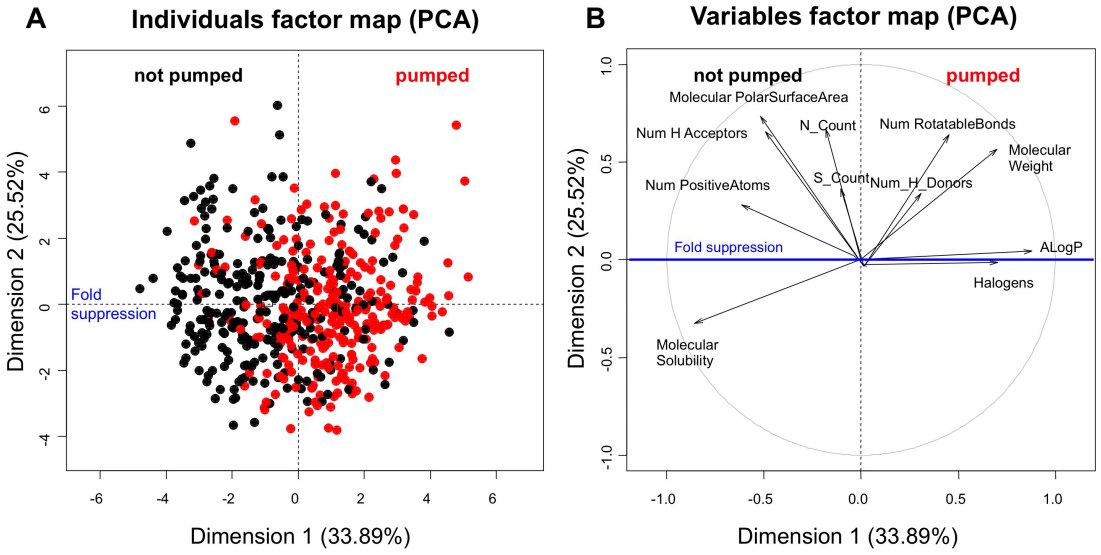


Figure 3.11: PCA analysis of AcrB pumped and non-pumped molecules showing data points and eigenvectors. **A** Molecules not pumped (black) and pumped (red) with the first two components explaining ∼60% of the data. **B** Fold suppression as quantitative variable is aligned with the x-axis. The directional eigenvectors for each parameter are depicted as arrows. Chemical properties used for the PCA analysis are HA, HD, RB, MW, number of halogens, NC, SC and PSA.

Descriptor	Pumped	Not pumped	p-value
Partition coefficient AlogP	$\geq$4.2	$\leq$2.8	1.20×10^{-13}
Solubility	$\leq$-5.9	$\geq$-4.3	1.20×10^{-13}
Molecular weight (g/mol)	$\geq$350	$\leq$288	1.20×10^{-13}
Number of positive atoms	0	$\geq$1	1.20×10^{-13}
Number of hydrogen acceptors	$\leq$3	$\geq$4	5.39×10^{-13}
Polar Surface Area	$\leq$69	$\geq$95	1.03×10^{-9}
Number of halogens (F^-, Cl^-, Br^-)	$\geq$2	0	8.16×10^{-9}
Number of rotatable bonds	4	4	0.8160

Table 3.1: Average values of the chemical descriptors that influence whether or not molecules are pumped by AcrB. The likelihood of a molecule being pumped or not pumped by AcrB increases or decreases according to the ranges displayed. Fold suppression is strongly influenced by properties such as molecular weight, solubility, polar surface area, number of hydrogen bond acceptors and number of halogen and positive atoms. Molecule flexibility represented as the number of rotatable bonds is indistinguishable between both classes. P-values were calculated using the Mann-Whitney test between two groups.

Further, we observed that benzoxazoles, thiophenes and triazines are less likely to be pumped, whereas carbazoles and anthracenes are frequently pumped. Interestingly, we noticed that molecules with functional nitro groups seem to reduce the effectiveness of molecules to be pumped as shown for MAC-0031858 and MAC-0031885 which differ in two nitro functional groups. Similar findings have been published relating to antimicrobial and antiparasitic drugs such as chloramphenicol, metronidazole and niridazole (Strauss 1979). The nitro group possesses a unique combination of chemical properties: it is strongly electron-withdrawing, small, increases the polar surface area, and can form hydrogen bonds. In addition, the nitro group can be bioactivated by enzymatic reduction to give reactive species. Several studies suggest that a nitroreductase-dependent mechanism catalyse those groups thus producing nitroso and hydroxylamine intermediates that can react with biomolecules exerting toxic and mutagenic effects (Patterson and Wyllie 2014).

To examine if there is a link between compound structure and potential non-reversible binding within the AcrB, we performed small molecule docking using available crystal structures from the Protein Data Bank (PDB). All the molecules in this study were docked using AutoDock[6] to the ligand binding site of three different crystal structures (PDB Identifiers: 2DRD, 2W1B, 1T9X) of AcrB protein on the basis of both the steric (or shape) complementarity as well as the interacting group complementarity and scored on the basis of their binding energy. The top 100 molecules from the three different docking results were analyzed further. 41 molecules which showed high docking scores for all three crystal structures were analysed to distinguish between interacting and non-interacting residues within each given protein. All three structures were crystallised with small molecules bound in the periplasmic domain considered the substrate binding pocket for the

[6]http://autodock.scripps.edu/

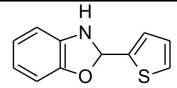
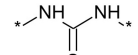
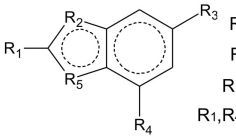
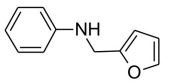
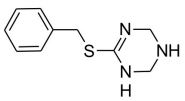
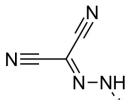
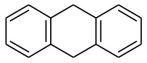
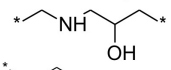
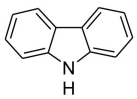
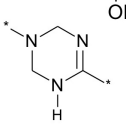
Murcko Assemblies	pumped	not pumped	Structural Features	pumped	not pumped
		5		28	3
			 R1,R3,R4=H; R5=N;R2=C*	11	1
			R4=H;R1,R3=C*;R2,R5=N	5	
	1	10	R1,R3,R4=H;R5=N;R2=C*	5	
			R1,R3,R4=H;R5=N;R2=SC*	5	
		11	R1,R4=H;R3= C*; R5=S,R2=C*	4	
				40	1
			 R = double bond	1	30
			R = pyran	7	90
	1	23			10
	3				12
	6(+12)	1			15

Figure 3.12: Bemis-Murcko assemblies and structural features connected to differential AcrB transport. Assemblies of the class benzoxazoles, anilines, thiophenes and triazines tend to not be pumped, whereas for carbazoles and anthracenes the likelihood of being pumped is high. Heterocycles containing indenes, indoles, thiazoles, benzothiophenes, benzimidazoles and isothiazolesare were likely to be pumped. Functional groups such as nitro, nitrile and triazine showed reduced transport by AcrB. Molecules that contained an ethylamino-2-propanol substructure were enriched in the not pumped class in the dataset.

active transport. Specific residues interacting (atoms within 6 from any atoms of the ligand) in case of actively pumped, not pumped and weakly pumped molecules were considered as potential interaction partners. It was observed that the substrate binding pocket is rich in aromatic amino-acid residues: Phe136 and Phe178, and Phe610, Phe615, Phe617 and Phe628. There are some polar residues in this area, such as Asn274 and Gln176, possibly forming hydrogen bonds with the drug molecule. Four proline residues, Pro718, Pro669, Pro565, and Pro833, surround the pocket.

Each residue suggested by AutoDock to form bonds with a molecule was counted and classified as pumped, weakly pumped and not pumped. Some of the residues were frequently considered as potential interaction partners with each of the classes. Surprisingly there were no residues discovered that showed a clear preference for either (weakly) pumped or not pumped for any of the three PDB structures. Similarly the docking score results showed only a weak correlation ($R^2 = 0.1$) for high ranked and low ranked molecules in relation with the logarithmic fold AcrB suppression ratio (2DRD, 1T9X).

In contrast, the NB learner and PCA showed quantifiable differences, with both methods being fundamentally different classifiers. The NB learner is a supervised classification algorithm that uses the density kernel estimates for the molecular descriptors and Bernoulli counts for the structural features to calculate a likelihood score that sums all feature probabilities. The PCA analysis is an unsupervised reductionist method to explain most of the variance with reduced dimensions that are ranked based on how much variance they can explain in the initial dataset. The descriptor space and structural features suggest that there are preferences that can be exploited to understand substrate specificity of MDR pumps. An interference with drug pumps would allow rational combination approaches to improve targeted drug efficiency. To overcome drug resistance, the combination with molecules that change the cell membrane or cell wall are considered the most efficient strategy. Supporting this argument are systematic drug combination studies

that tend to discover synergistic molecule that act on the permeability of cells and lead to potentiated drug action. A selective drug pump inhibitor would provide a targeted approach to reduce the pathogenic load in multicellular organisms.

3.2.2 Small molecules that mediate neuronal stem cell fate

Together with Phedias Diamandis, a graduate student in the lab of Peter Dirks, we screened the LOPAC for mediators of neuronal stem cell fate (Diamandis et al. 2007). The signalling pathways that govern proliferation and self-renewal of Neural Stem Cells (NSCs) were largely unknown when we conducted this chemical-genetic screen to probe the operational circuitry of NSCs. For the screen, primary mouse embryonic neurospheres were grown for 7 days and then harvested, enzymatically digested and mechanically separated. Viable cells were plated at low densities in 96 microwell plates. Compounds were added at days 1 and 4 at a final concentration of 3 μM. After seven days, viability of cells was assessed by quantifying 3-[4,5-dimethylthiazol-2-yl]-2,5-diphenyltetrazolium bromide (MTT) vital dye. The screen showed plate-to-plate variability and edge effects due to evaporation of wells, which I was able to correct computationally with appropriate normalisation procedures (Brideau et al. 2003). We identified chemical scaffolds associated with molecules that acted on a variety of signalling pathways. For example, we confirmed phenothiazines and apomorphines as regulators of cell differentiation in murine embryonic NSCs. Further, this study identified potential anti-proliferative agents against central nervous system tumours, and showed that most neurotransmission pathways in the mature central nervous system also influence neuronal stem cell fate (Diamandis et al. 2007). The ability to modulate stem cell differentiation patterns has profound impact on stem cell biology and will lead to advances in clinical applications such as tissue

regeneration and cancer treatment. The hypothesis that the majority of cancers involve stem cells, including progenitor cells, has become established in recent years (Fábián, Vereb, and Szöllősi 2013; P. A. Jones and Baylin 2007; Tan et al. 2006). Our study focused on brain cancer cells, but our research has been cited by clinicians and researchers working on breast, bladder, renal and colon stem cell cancer subtypes (McDermott and Wicha 2010). Previously, the scientific literature was mainly focused on developmental signalling pathways such as the Notch and Wnt pathways, which were thought to be the major mediators of resistance in cancers (Ranganathan, Weaver, and Capobianco 2011; Woodward et al. 2007). However, this study and others (Hanahan 2014) have shown that serotonergic drugs prevent cancer cell growth. In addition, it has been suggested that serotonin itself could be used as a biomarker for cancer detection in urinary bladder (Siddiqui et al. 2006; Soll, Jang, et al. 2010; Soll, Riener, et al. 2012), adenocarcinoma and renal cell carcinoma (McDermott and Wicha 2010). Current models suggest that serotonin receptors are expressed on a wide range of cell types that either function as mediators or have regenerative function in heterogenic cell assemblies (Nocito et al. 2008; Soll, Riener, et al. 2012). Our work also motivated systematic studies in drug repurposing using FDA approved drugs (Pollard et al. 2009; Sachlos et al. 2012) and *in vivo* stem cell tumor screens (Caussinus and Gonzalez 2005). Supporting our findings, a screen with 30,000 synthetic compounds found four dopamine derivatives that inhibit growth of glioma-derived neural cell lines (Visnyei et al. 2011). Similar findings have been reported for peripheral dopamine; it influences metabolism and growth in tumors (Rubí and Maechler 2013; Scemama 1984). The importance of this work was the insight that signalling pathways previously thought to only operate in the mature Central Nervous System (CNS) actually dictate neuronal stem cell survival, and by extension brain cancer stem cell proliferation and survival.

3.2.3 Bioavailability of small molecules in worm

Parasitic nematodes infect 24% of the world's population and are widespread in crops and livestock according to the World Health Organization. Treating parasitic worms is costly and labour intensive and target-based approaches have failed to yield novel anthelmintics. Bioavailability is a prerequisite to bioactivity and the soil-dwelling nematode *C. elegans* is a useful model organism to select such features since it has extensive enzymatic and physical protective barriers. The worm's genome is filled with predicted xenobiotic detoxification enzymes, including 86 cytochrome P450s and 60 ABC transporters that are likely to function as efflux pumps. In comparison, the human genome encodes 57 cytochrome P450s and a total of 48 ABC transporters (S.-F. Zhou, J.-P. Liu, and Chowbay 2009). The hit rate of small-molecule screens can be increased significantly by pre-selection of molecules that have an increased likelihood of reaching their target in the nematode. In collaboration with the Roy laboratory I assessed accumulation and metabolism of 1000 drug-like small molecules in *C. elegans*. A reverse phase High Performance Liquid Chromatography (HPLC) coupled with Diode Array Detector (DAD) was used to separate and detect absorbed and metabolised compounds in the worm lysate. Similar approaches are used in clinical and forensic studies (Herzler, Herre, and Pragst 2003) and for the identification of novel natural products in microbial, maritime and plant extracts (Wolfender 2009). The initial dataset contained a small sample of molecules of antipsychotic drugs used to showcase separation using PCA based on chemical properties 3.13. An extended analysis based on PCA highlighted the unsuitability of this approach to distinguish between processed, retained and metabolised compounds.

First we identified 1,132 compounds that are detectable by HPLC at 7.5 nM. These compounds were from the Spectrum collection of 2,000 pharmacologically

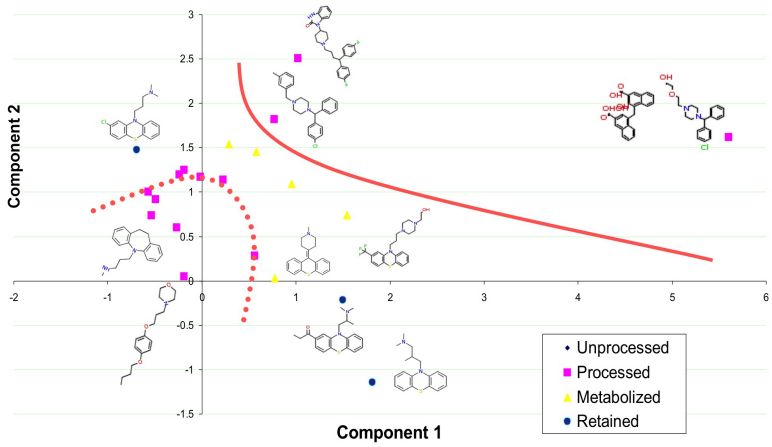


Figure 3.13: Initial compound set to develop a worm retention model. Molecules are 'placed' based on their chemical properties used for the PCA analysis (HA, HD, RB, MW, number of halogens, NC, SC and PSA). Molecules that are visible on the HPLC-DAD (processed) can either be retained or metabolised leading to changes of the HPLC signal.

active compounds as well as known worm-active compounds. These 1,132 compounds were screened for accumulation in wild-type worms after a 6 hour incubation period. For each compound, the HPLC spectrum of compound alone was compared to the spectrum of compound-treated worm lysate. Further, to distinguish between metabolised compounds and metabolites present in the lysate, compounds were tested for drug precipitation, compound sticking to worm cuticle (Sodium Dodecyl Sulfate (SDS) wash controls) and dead worm controls. Only 483 compounds passed a detection limit test on the HPLC-DAD and 74 compounds were identified as accumulating in worms. I applied machine learning to Enhanced Connectivity FingerPrints of length 4 (ECFP4) and Murcko fragment (Bemis and Murcko 1996) representations of the 483 compounds to identify characteristic structural features for accumulating and non-accumulating compounds. We discovered that 72 distinct compounds are associated with three chemical scaffolds, which account for 41% of the accumulating compounds (Figure 3.14). The structural scaffolds I identified, notably the biaryl scaffold, the piperazine

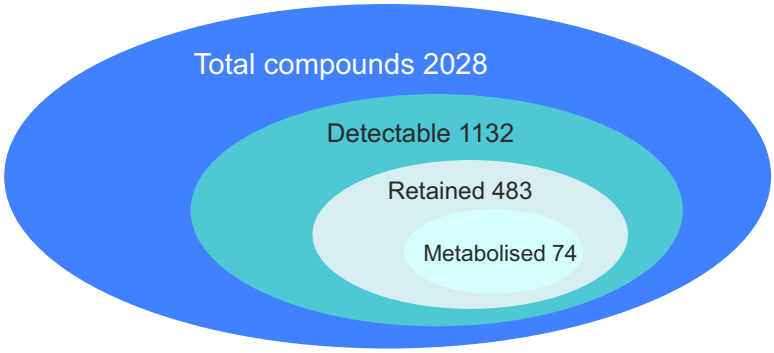


Figure 3.14: Molecule refinement layers of the worm bioaccumulation model. Out of 2,028 compounds tested, 1,132 were detectable on the HPLC, 483 produced a signal above the detection limit for retention of which 74 accumulated or metabolised in worm.

scaffold and the benzopyran substructure in the 2- or 3-phenyl-chromen-4-one scaffolds have all been previously identified as privileged substructures (Y. Chen and Shoichet 2009; Horton, Bourne, and Smythe 2003; Klekota and F. P. Roth 2008). A privileged substructure is defined as a single molecular framework able to provide ligands for diverse receptors (Evans et al. 2002). Specificity can be achieved by varying the substituents that decorate the privileged scaffold (Mason et al. 1999). For example, the biphenyl scaffold, which accounts for one-third of the accumulating unfused biaryls in worms, is found in 4.3% of all known drugs, representing molecules from diverse therapeutic classes. Statistical analysis of Nuclear Magnetic Resonance (NMR)-derived binding data for 10,080 compounds (represented by 104 structural fragments) and 11 protein targets identified the biphenyl scaffold as a privileged substructure that preferentially binds proteins (Hajduk et al. 2000). In the worm accumulation model almost 40% of biaryl scaffold compounds accumulate as metabolites. However, it is not possible to identify the active form of a molecule that accumulates since compounds are often metabolised to their bioactive form. For any given bioactive compound discovered through an *in vivo* screen, it is thus not certain that the parent structure is the bioactive species. A recent follow up study from the Roy laboratory

returned its focus on a lethality phenotype (Burns, Luciani, et al. 2015). In retrospect, the dataset built with the bioaccumulation model did not include a fitness assessment and only focused on compound absorption in worms. The enrichment of retained and metabolised compound fractions indicated a very small compound space that was dominated by three core scaffolds. This is especially critical when considering the composition of the structurally diverse Spectrum compound library. Only 1,132 molecules contained detectable chromophores on the HPLC-DAD system. Of those 1,132 molecules, only 483 molecules passed another inherent detection-limit test, which biased the set towards compounds that are visible below an approximate concentration of 19 μM. Only 74 of these 483 compounds (15.3 % in the refined dataset) accumulated in worms. Approaching the final number from a conditional probabilistic viewpoint, the overall true effect size of a generalistic bioaccumulation model is likely to be low. If in this study we estimate an optimistic 80 % power (small β-error), an alpha error of 5 % and a detection bias odds ratio $R = 483/2100 \approx 23\%$, then the Positive Predictive Value (PPV): $PPV = 0.8 * 0.23/(0.8 * 0.23 + 0.05) = 0.184/0.234 = 0.79$; that is, 79 % of our results are expected to be correct. Now focused only on the 3.5 % compounds that accumulated in our initial dataset the power of the study drops down to 36 % PPV. Given the size of the compound library, it is obvious that our experimental approach introduces a bias into the dataset by selecting for HPLC-detectable compounds. It would be possible to pre-screen focused chemical libraries for bioaccumulation studies *in silico* to identify HPLC-detectable compounds with potential chromophores (Masunov and Mikhailov 2010; Mizera et al. 2015; Stanstrup, Neumann, and Vrhovšek 2015).

In summary the experimental setup in our study was tailored to provide a fast bioaccumulation snapshot. The initial expectation was that many compounds are HPLC-traceable and will be absorbed through the intestine and therefore

provide sufficient data for identification of scaffolds and moieties for future nematicides. We expected that the bioaccumulation model will help explain differences in outcome between *in vitro* and *in vivo* screens. A similar study performed in *Haemonchus contortus* to assess the bioaccumulation in a parasitic nematode using Liquid Chromatography-Mass Spectrometry (LC-MS) data, suggests that a partial least squares regression model using four physicochemical properties (CLogP, HD, RB and E-state) has predictive power for bioaccumulation (X. Zhou et al. 2014). Collectively, this work illustrates the complexity and experimental difficulties encountered when generating robust and unbiased datasets for computational predictive models. The limitations of our bioaccumulation model (Burns, Wallace, et al. 2010) might be overcome through additional models that detect the presence of chromophores, predict retention times and likelihoods of metabolic degradation. Such models in combination with a phenotypic prediction model built from large phenotype studies could guide discovery of novel nematicides.

3.3 Application of cheminformatics to chemical-genetic data sets in yeast

Explorative mode of action studies constitute an important aspect of chemical biology. Functional genomics approaches in the budding yeast *Saccharomyces cerevisiae* offer a powerful platform to assess and compare the effects of small molecules (Ericson et al. 2008; Giaever et al. 2004; Hillenmeyer et al. 2008; Parsons, Brost, et al. 2003; Winzeler et al. 1999). In early studies of yeast chemical-genetic interactions on a genome-wide scale, chemical properties played only a minor role in the assessment of compound mode of action. A collaboration with the Giaever lab allowed me to integrate functional genetics with cheminformatics

to improve interpretation of chemical-genetic profiles of small molecules (Ericson et al. 2008; Hillenmeyer et al. 2008).

3.3.1 Analysis of chemical-genetic data

A pool of ∼5000 deletion strains can be grown in the presence of small molecules to obtain quantitative fitness scores that are indicative of how a particular strain responds to the compound. Growth advantage or disadvantage of a specific deletion strain represent chemical-genetic interactions. The Giaever and Nislow laboratories generated chemical-genetic profiles for a compendium of ∼1400 experimental conditions to understand how yeast responds to a wide range of external stresses. The stress conditions included natural products, approved drugs, heavy metals and salts as well as compounds with unknown mode of action. The HaploInsufficiency Profiling (HIP)/HOmozygous deletion Profiling (HOP) dataset contained 333 molecules compared with previously published atasets of 12 and 82 compounds (Parsons, Brost, et al. 2003; Parsons, Lopez, et al. 2006). To obtain comparable results between the molecule responses, growth of the deletion pool was calibrated to an approximate 10% inhibition compared to solvent-treated comtrol samples. To illustrate the complexity of chemical-genetic profiles, Figure 3.15 highlights the occurrence of outliers for different compounds, compared with outliers for different genes.

To reduce the data matrix to significant hits for genes and compounds, I applied an outlier selection procedure based on boxplot statistics in two dimensions, across genes and across compounds. If a gene y was an outlier in response to a specific compound a and compound a is an outlier among all compounds for gene y, then the combination of y-a gained a score of two (Figure 3.16A). The figure also illustrates that such discretisation helps to highlight differences between

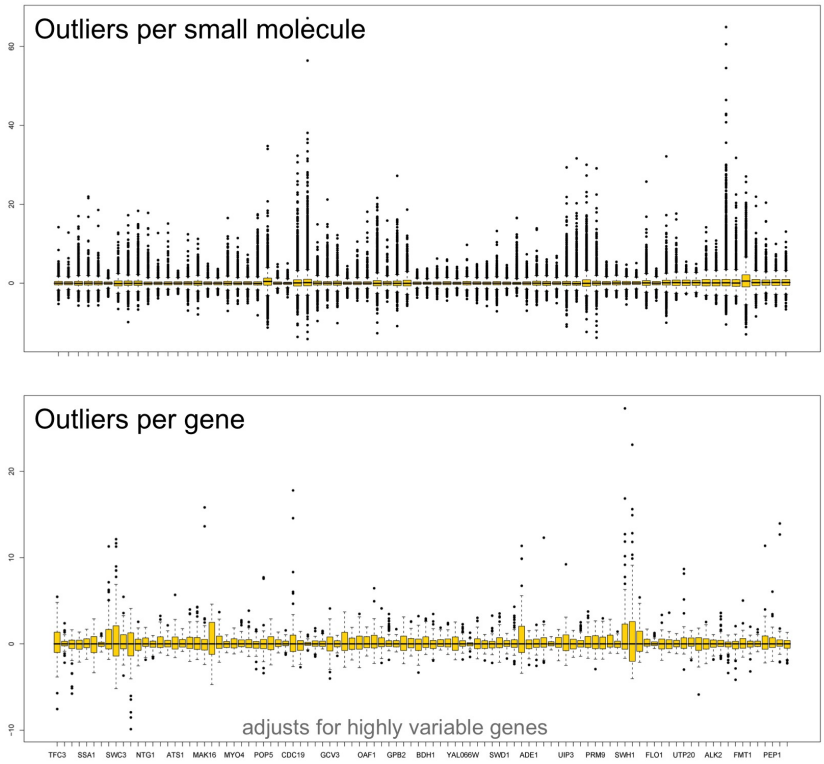


Figure 3.15: Two dimensional analysis of chemical-genetic data sets. Top: Median normalised compound profiles shown as boxplot for 74 molecules. Bottom: Median normalised gene outliers shown as boxplot for 74 genes. Outliers beyond the hinges are indicated as dots.

cluster members. Cluster 30 was identified using average linkage hierarchical Tanimoto similarity clustering and significance testing (Park et al. 2009). All four compounds were structurally similar (Figure 3.16B). For three of the compounds we get Gene Ontology (GO) enrichment for the Tryptophan (Trp) synthesis pathway. Interestingly, RDR 03963 predicted to be an inhibitor of *BNA2*, a tryptophan 2,3-dioxygenase, does not show sensitivity to any of the genes in the pathway (*ARO1,2* and *TRP1-5*) (Figure 3.16C). This is not necessarily surprising as we have shown for homozygous genes such as CNB1 and TOP2 that in presence of an inhibitor, those strains produce a wildtype phenotype. Inhibition of *BNA2* will prevent Trp degradation and is therefore likely to deactivate a flux controlled production pathway preventing a growth phenotype (Miozzari, Niederberger, and Hütter 1978). The compounds RDR 03963, SPB 05768 and SEW 05400 produce a hookless phenotype with a short root in *A. thaliana*, associated with the auxin biosynthetic production in plants that rely on Trp as initial building block (Stepanova et al. 2011). The plant phytohormone, Indole-3-Acetic Acid (IAA), is synthesised by a Trp dependent and independent pathway. Two studies using single deletions (ibid.), and double deletions of genes between the Trp and the parallel pathways show that the IAA is synthesised via the Trp dependent IAA biosynthesis pathway at the root tip (Kiyohara et al. 2011), producing the observed phenotype.

Further, figure 3.17 highlights a few complex examples where the 2D boxplot improved selection of significant deletion strains for follow up on compounds and Mode Of Action (MOA). To identify significant outliers between two screens, we took the sum from both boxplot discrete count vectors, that now range from zero (in IQR per compound and gene) to absolute four (outlier per compound and gene). This simple selection method provides three confidence levels, colour coded in Figure 3.17A-E for each compound replicate pair or pairwise comparision. The

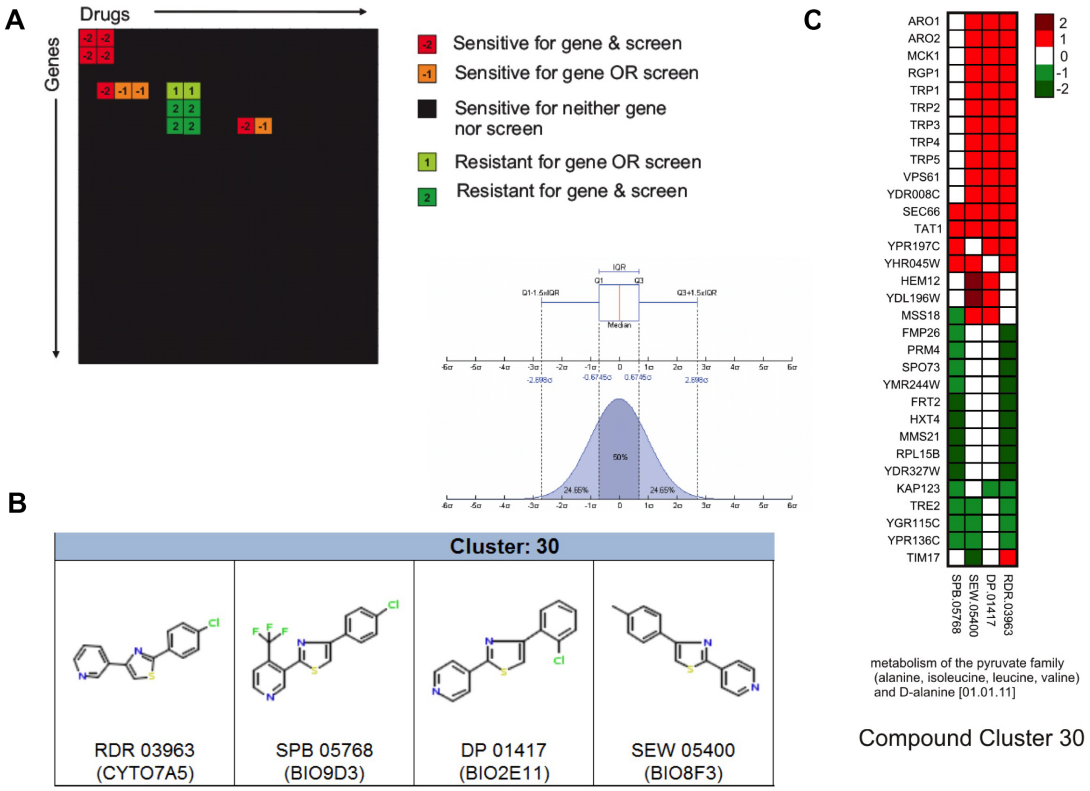


Figure 3.16: Selection of hits in chemical-genetic yeast screen. **A** Visualization of the 2D boxplot hit selection. **B** Structures in Tanimoto similarity cluster 30. **C** Chemical-genetic cluster with at least two sensitive genes across four compounds RDR 03963, SPB 05768, DP 01417 and SEW 05400.

first four scatterplots show biological replicates for camptothecin, acriflavinium, clotrizamole and amphotericin B (Figure 3.17A-D), with different R^2 values. Ideally the MOA of a molecule is revealed with few significant deletion strains providing a clean chemo-genomic profile (Figure 3.17A). The opposite is likely with molecules that are promiscuous binders and interfere in many different biological processes with less significant outliers (Figure 3.17B). Further, different compound concentrations affect the selectivity and replicability and can shift deletion strains into a phenotypic response (Figure 3.17C). Another possible scenario highlights compounds were the MOA is known but phenotypic response does not highlight deletion strains in close physical or functional proximity (Figure 3.17D). The polyene amphotericin B was the first widely used antifungal drug introduced in the 1950s. It penetrates the fungal membrane and binds to ergosterol, leading to membrane damage. The chemo-genomic GO enrichment response highlights severe cell stress, chromatin modification and changes in metabolic processes and transcriptional regulation, also shown by a metabolic study (Belenky, Camacho, and J. J. Collins 2013). Figure 3.17E showcases the response for doxorubicin and daunorubicin. Both molecules are similar anthracyclines and clinically used for the treatment of cancer. The presumable MOA of doxorubicin is defined by a specific intercalation with the DNA double helix. But the chemo-genomic GO enrichment in yeast highlights genes needed for the integrity of the nuclear membrane (p-value 1.887×10^{-5}). Similar, the disruptive effect of doxorubicin on the nuclear membrane has been shown in a hepatocarcinoma cell line (Huh-7) (Eom et al. 2005).

Systematic gene deletions that render cells inviable have been studied extensively in haploid yeast. The large number of non-essential genes still puzzles scientists and has shaped the idea of parallel pathways. To uncover these buffering mechanisms, systematic double deletion screens have been pioneered by the Boone lab and others to map genetic interactions in the cell.

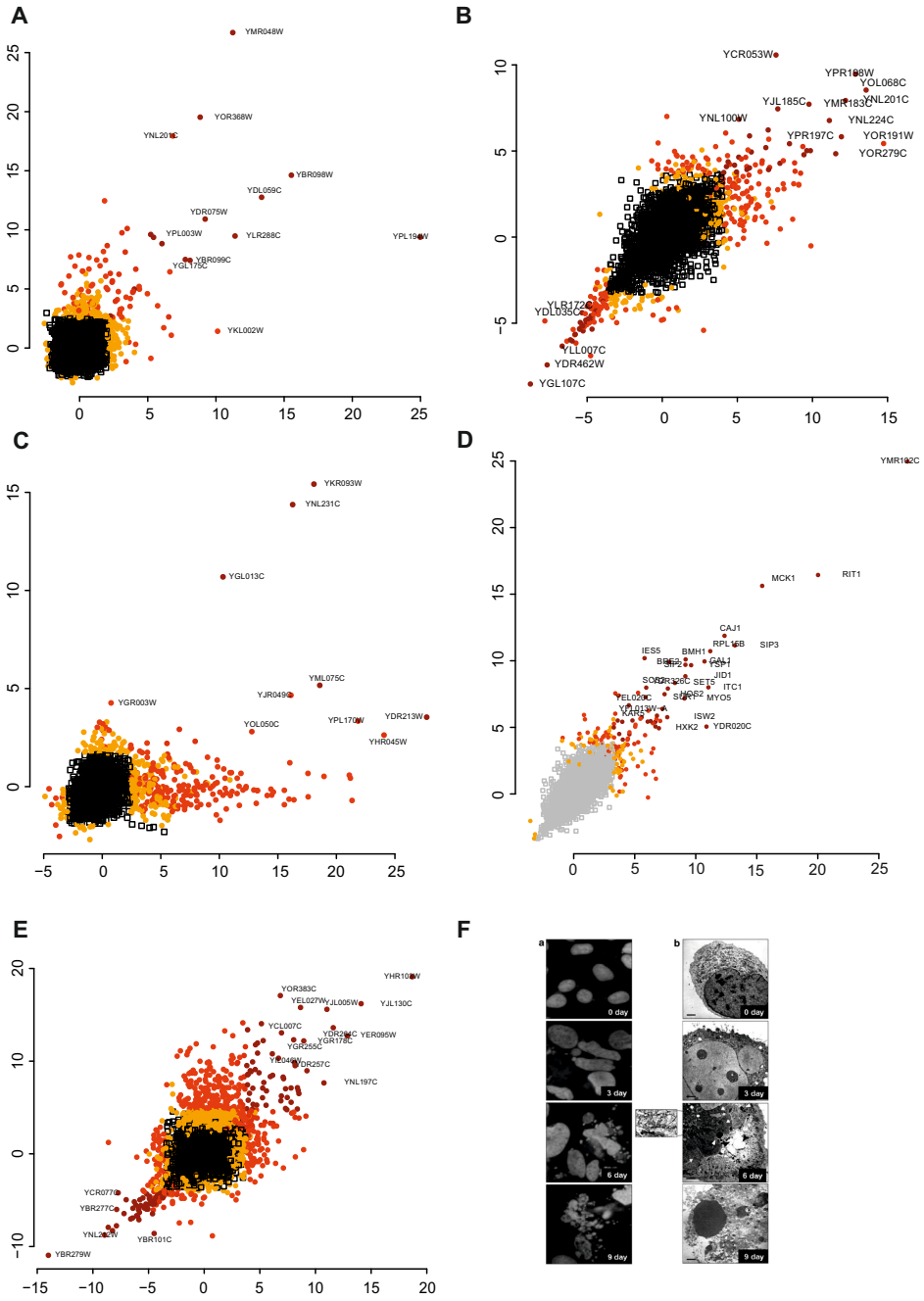


Figure 3.17: Chemical-genetic profile scatterplots for natural products. **A** Two replicates of 10μM camptothecin ($R^2 = 0.06$) **B** Two replicates of 40μM acriflavinium ($R^2 = 0.51$) **C** 47μM and 82μM of clotrizamole ($R^2 = 0.18$) **D** Two replicates of 62μM amphotericin B ($R^2 = 0.68$) **E** 47μM of daunorubicin (x-axis) and 82μM of doxorubicin ($R^2 = 0.17$) **F** Microscopic images of Huh-7 cells treated with 50 ng/ml doxorubicin show collapse of nuclear membrane (left: fluorescent stain with Hoechst 33258; right: Electron microscope; top to bottom 0, 3, 6, 9 days; taken from (Eom et al. 2005)). Significance score: $\geq |3|$ (brickred), $|2|$ (brightred), $|1|$ (orange).

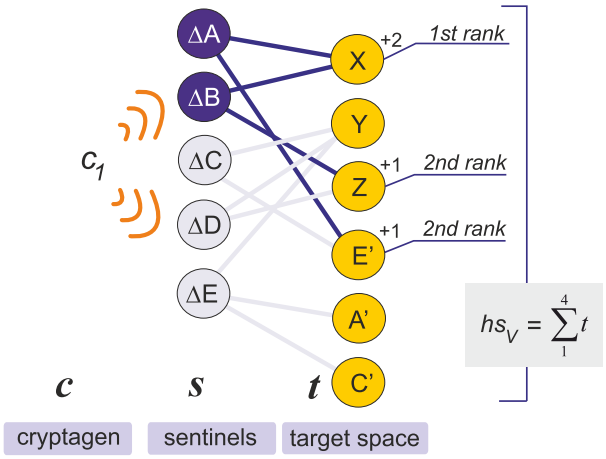


Figure 3.18: Exploration of chemical-genetic data with bipartite graphs. Compound c_i induces genetic response of deletion strains s. The target space t is derived based on synthetic genetic interaction data from BioGRID.

The complexity of the indirect genetic response to small molecules has been a major challenge in assigning MOA to novel chemical matter. To address the nature of the indirect response, the known interaction network created by the scientific community can be combined with the small molecule response to study their relationship by using bipartite graphs. This Second Order Neighbour Activity Response (SONAR) method is explained in Figure 3.18 (Wildenhain, Spitzer, Dolma, et al. 2015). The algorithm first builds a bipartite graph to narrow in on the most likely targets of a single compounds c_i, and then ranks the genetic interaction edges based on the interaction between s and t. The target space and target genes for single compounds were identified through the following steps:

1. Sensitive deletion strains were selected based on the 2D boxplot with a significance score ≥ 2.

2. Genetic interactions for the set of sensitive deletion strains V_S were used to identify all neighbours V_T that formed the target space. V_S and V_T are sets of nodes of a bipartite graph that are connected by an edge

if a genetic interaction between two genes has been reported (based on BioGRID release number 3.076) (Chatr-aryamontri et al. 2015). No edge in E connects vertices within the same set of nodes. The graph is defined as $G = \langle V_S \cup V_T, E \rangle$, where $V_S = \{s_i | 1 \leq i \leq k\}$ and $V_T = \{t_j | 1 \leq j \leq n\}$. To represent edges within V_S, we introduce the connected nodes into V_T represented as $V_T = \{s_i^{'} \wedge t_j | 1 \leq i \leq m, 1 \leq j \leq n\}$. Each s_i is either sensitive $s_i^+ = 1$ or non-active $s_i^- = 0$. The weight of edge E_{ij} for a pair of nodes s_i, t_j is defined as $s_i * m$ where m is the number of inferred interactions between these nodes based on genetic interaction data. To rank the nodes in V_T, each node t_j is assigned a score that is the sum of the weights of all edges that link to t_j.

3. The sum of the $n = 35$ highest t_j ($hs_V = \sum_1^n t_j$) is used to characterise the target space of each compound. P-values are calculated from 1,000 permutations to estimate a background distribution using as many s_i^- as there are s_i^+ (Spitzer, Griffiths, et al. 2011; Dittmar et al. 2013). The value of n was chosen empirically to balance target space size with computational costs.

4. If there is no significant enrichment in hs_V over the background permutations (p-value ≥ 0.05), return to step 2, remove the weakest bioactive hit and re-run analysis. Continue to loop through steps 2-4 until SONAR hs_V is significant or terminate calculation if total number of s_i^+ is reduced to 4. If a significant gene set s_i^+ is found, this defines target space set t_i^s.

5. The highest ranked genes in t_i^s are most likely to represent actual targets of molecule c_i.

3.3.2 Integrating cheminformatics with yeast-chemical genetic data

For the Hillenmeyer et al. study, I integrated a chemical properties and structures of the bioactive molecules with the biological responses elicited by each compound. Findings from this analysis included strong correlations between the impact of protein complex formation with planar tricyclic structures and the toxic effect of amines on deletion strains in the tryptophan biosynthesis pathway. A key finding of this study was that 97% of all yeast gene deletion strains can be specifically perturbed by at least one small molecule, and that transport process related gene deletions tend to make cells highly susceptible to small molecules. This work highlights the complexity of chemical responses in a single cell organism (Hillenmeyer et al. 2008), and has resulted in 600 citations to date. To extend the translational value to human health, I initiated another project that assessed the effects of psychoactive drugs in budding yeast (Ericson et al. 2008). Psychoactives have been dubbed dirty drugs because of their often severe side effects. The yeast deletion strain collection proved an ideal testing ground to explore physiological responses to this class of compounds. Almost 60% of all deletion strains that have conserved homologs in human are sensitive to dopaminergic and serotoninergic drugs. Hydrophobicity is a key determinant of such effects since psychoactive drugs with an ALogP value below three are unlikely to show any activity against budding yeast, excluding more than half all molecules from the study. For the remaining drugs I showed that compounds influence aromatic amino acid biosynthesis, chromatin remodelling, vesicle mediated transport and protein localisation. This work emphasised budding yeast as model for pharmacogenetic studies to investigate off-target effects of clinically relevant psychoactive drugs (ibid.). A recent paper extended the previous efforts by testing 3,250 molecules from a synthetic library to find chemical substructures that could be linked to 4,500 deletion strain sensitivity responses (Lee et al. 2014). The authors of this study concluded

that a core set of 317 molecules and 121 genes represent a consistent activity relationship in the data. Considering the size of the initial dataset (4,500 deletion strains, 3250 molecules), the results suggest that either the functional classification of genes (Gene Ontology) and the set of chemical substructures (rings and ring systems) or the data analysis used in this study might limit interpretation of the data.

In summary, the analysis of extensive chemical-genetic profiles in yeast enabled the characterisation of the mode of action of small molecules and their combinations, as well as off target effects. Except for very few cases, it has proven difficult to determine the exact mode of action of small molecules. For example, amphotericin B has been used as an antifungal drug for decades and it was only discovered last year that it extracts ergosterol from fungal membranes and sequesters it in extracellular sponges. Amphotericin B had been known to bind ergosterol, but it was assumed that the amphotericin B-ergosterol complexes form pores in the membrane. Similarly, many other compounds known to integrate into cellular membranes also affect a variety of cellular processes. Cationic amphiphilic drugs, for example, affect cellular membranes and most of these molecules have a characteristic chemical-genetic signature that includes the homozygous deletion strains SLT2, BCK1 and DRS2. However, it is hard to narrow in on the exact mode of action because the effects of very few compounds are characterised in detail. A further key point is that most compounds do not have a single protein target in the cell but instead modulate multiple proteins. The concept of polypharmacology (Besnard, Ruda, et al. 2012; Dar et al. 2012; A. L. Hopkins 2008) has resulted in a fundamental change in how we approach drug discovery. Current approaches to address the problem of multiple targets focus on the integration of network biology (Nelander et al. 2008) and the application of machine learning to the analysis of large-scale data sets (Stephan, Stegle, and Beyer 2015; Wildenhain, Spitzer, Dolma, et al. 2015; Xiong et al. 2015). Previous work

on modelling biological network behaviour has indicated that the buffering effects in interaction networks can be robust to selective perturbations, with the implication that combinations of small molecules may be much more effective than single agents (Wildenhain and Crampin 2006; Kitano 2007; Lehár, Zimmermann, et al. 2007; Lehár, A. Krueger, et al. 2008). A systems-level approach based on genetic and protein interaction networks has been suggested as a way to prioritise potential antimicrobial drug targets (Roemer and Boone 2013). To test such hypotheses, a graph based algorithm that uses chemical genetic interaction data to predict small molecule drug combinations was developed. The key component of this algorithm uses two machine learning steps, first to find significant compound features that are associated with deletion strain responses, and second, to learn deletion strain responses that are associated with synergistic activity responses. Verification of predictions from the algorithm lead to novel drug combinations that have been tested against pathogenic fungi and Human Embryonic Kidney (HEK) and HeLa cells (Wildenhain, Spitzer, Dolma, et al. 2015). The concept of using genetic interaction data to suggest drug combinations fostered an informatics project (Winter, Wildenhain, and Tyers 2011) to develop a REpresentational State Transfer (REST) that allows facile access to interaction data from our in-house BioGrid database (Chatr-aryamontri et al. 2015).

Chapter 4

MolClass: a web portal to interrogate diverse small molecule datasets with different computational models

Commercial tools, such as Pipeline Pilot, have been successfully used to build predictive models for drug discovery (Besnard, Ruda, et al. 2012; Burns, Wallace, et al. 2010; Keiser, Setola, et al. 2009; Rogers, R. D. Brown, and Hahn 2005; Clark, Dole, et al. 2015; Xia et al. 2004). This chapterintroduces MolClass, a novel open source cheminformatics tool that enables users to perform molecule activity predictions (Wildenhain, Fitzgerald, and Tyers 2012). MolClass uses a wide range of machine learning algorithms and compound structure similarity and property calculations to interrogate small molecule mechanism of action. MolClass currently contains data on more than 200,000 unique molecules with predictions for 18 pharmacological models, 40 screens and 6 compound libraries. For new compounds, MolClass calculates likelihood scores for each

of the pre-existing or user-generated models. MolClass provides a framework to verify structure-activity relationships within a user-defined dataset, and suggests molecules from the database as potential new scaffolds. This resource thus enables the biomedical researcher to interrogate screen hits in an efficient and user-friendly manner without the need for expertise in cheminformatics, machine learning or statistics. MolClass is also a resource that hosts reference datasets and provides predictive reference scores (Cheng, W. Li, et al. 2012; Clark, Dole, et al. 2015).

4.1 Data sources for MolClass

The main aim of MolClass was to build a tool that integrates existing activity information with open source cheminformatics software to guide hit selection for follow up experiments from screening data. In the past 10 years, the size of chemical data repositories has grown steadily, enriching the public knowledge of compound-bioactivity relationships. The major repositories for small molecule screen data are the National Institute of Health (NIH) PubChem database (B. Chen, Wild, and Guha 2009; Y. Wang, Bolton, et al. 2010; Y. Wang, Xiao, et al. 2012), ChEMBL, the European data resource for small molecule binding affinities (Fechner et al. 2013; Gaulton et al. 2012), the curated pharmacological interaction database Drugbank (Law et al. 2014; Wishart 2008) and a screening database hosted by the Broad institute called ChemBank (Seiler et al. 2008; Petri Seiler et al. 2011). Compound-activity relationships are scattered over a vast set of different *in vivo* and *in vitro* screens. Notably, a small number of compound entries in PubChem are associated with hundreds of screens while the majority of compounds currently have no biological activity data deposited. Further, some High Throughput Screen (HTS) datasets have detailed dose-dependent affinity readouts while others are primary screens that simply provide a qualitative classification into active and non-active compounds. ChEMBL,

due to its systematic annotation of drug target relationships and continuous development provides a perfect resource to evaluate established drug target and compound spaces and their interconnectivity (Heikamp and Bajorath 2011; A. L. Hopkins 2009). ChEMBL has been used successfully to predict new targets for old drugs and old targets for new compounds (Keiser, Setola, et al. 2009). Most compellingly, it has become a popular data source for building classifiers for small molecule target prediction (Afzal et al. 2015; Clark and Ekins 2015; Koutsoukas et al. 2013) and therefore replaced WOrld of Molecular BioAcTivity (WOMBAT) (Olah et al. 2008) and Molecular Drug Data Report (MDDR) (BIOVIA and Thomson Reuters) as first choices for benchmarking (Southan, Várkonyi, and Muresan 2009; Nidhi et al. 2006). To provide additional information for small molecule screens performed by collaborators and the lab, we used a multinomial Naive Bayes (NB) classifier (Rogers, R. D. Brown, and Hahn 2005) to predict putative targets derived from a recent ChEMBL release. Curated information is available in DrugBank, a focused resource on pharmacologically active compounds and approved drugs (Law et al. 2014). It provides a good reference to evaluate models and propose annotated drug reference points for novel molecules. ChemBank presents a large screening repository (Seiler et al. 2008; Wagner and Clemons 2009) with a structure similar to our ChemGRID in-house database, and therefore can serve as datasource for MolClass. Among other applications, ChemBank data was used to build a predictive model to identify potential chromatophores as false positives in screens with fluorescence readouts. Those resources provide existing activity data for small molecules and consequently allow to build predictive compound target models for discovery of compound mode of action. MolClass was built with the purposes of facile management and utilisation of our in house screening data, and to cross-validate experimental outcomes with other datasets.

4.2 Machine learning in MolClass

MolClass was designed to provide a wide range of machine learning classifiers to test and compare model performance across different datasets. Popular algorithms for learning relationships are NB, Random Forest (RF) (Breiman 2001), Support Vector Machine (SVM) (Cortes and Vapnik 1995) and logistic regression (Cox 1958). In the recent bioinformatics literature, two approaches have dominated applied Machine learning (ML) algorithms: SVMs that use either linear or polynomial kernels and RFs (Schierz 2009; H. Yu et al. 2012; Balfer et al. 2014). RFs have become popular due to the ease of parameter choice and the low risk of overfitting (Statnikov, L. Wang, and Aliferis 2008; Cortes-Ciriano, Bender, and Malliavin 2015). Support vector machines have the advantage of different kernel functions that, in addition to expert knowledge, can utilise complex segregation functions, as opposed to random forests that use repetitive hierarchical separation throughout the dataset. A wide range of other learners are available in MolClass including: K-Nearest Neighbours (KNN) (Fix and Hodges Jr. 1989) to establish a base level predictor as reference point; NB (Maron 1961) as an established molecule prediction algorithm (Besnard, Ruda, et al. 2012; Feng et al. 2009; Glick et al. 2004; Keiser, B. L. Roth, et al. 2007; Koutsoukas et al. 2013; Rogers, R. D. Brown, and Hahn 2005); LogitBoost (J. Friedman, Hastie, and Tibshirani 2000) as a representative algorithm for logistic regression; J48, the classical decision tree algorithm (Quinlan 1986); Logistic Model Tree (LMT) (Landwehr, M. Hall, and Frank 2005) and NB tree learners. Further, MolClass employs BayesNet (N. Friedman, Geiger, and Goldszmidt 1997) as a local search algorithm and a feed forward Multilayer perceptron algorithm (Reed and Marks 1998) for molecule classification. Based on the literature, best performances are expected from a RF classifier, followed by SVM (Fernández-Delgado, Cernadas, and Barro 2014). Similar results for machine learning algorithms have been shown in applied cheminformatics and drug discovery (Schierz 2009; H. Yu et al. 2012). Generally,

it is prudent to apply several methods since each method selects different subsets of actives (Sheridan and Kearsley 2002). Comparing the performance within a multi method ensemble learner in MolClass, assessed by a weighted regression algorithm, showed that the best performances were delivered by a RF classifier (6 out of 7 datasets assessed). The Area Under the Curve (AUC) for each approach for standard datasets was determined as RF, 0.896; Ensemble, 0.895; BayesNet, 0.878, NBTree, 0.843; KNN, 0.821; LogitBoost, 0.82; LMT, 0.818; SMO: 0.804; J48, 0.802; LibSVM, 0.784; NaiveBayes, 0.766 (see Table 4.1). Given the excellent performance of Bayesian networks, it is surprising that this approach is rarely used in cheminformatics applications. In contrast, support vector machines (SMO and LibSVM) are a popular choice, but did not perform well in MolClass, even with iterative parameter grid search optimisation. Since the random forest learner outperforms all other algorithms, it is considered as baseline reference model in MolClass.

4.3 Small molecule descriptors in MolClass

It is important to note that the choice of the chemical descriptors impacts performance of any algorithm. Two major types of descriptors are usually emphasized, physical property descriptors such as polarity, charge, torsions, weight and solubility and structural descriptors that represent discrete chemical sub-features. The structural fingerprints are either built by an iterative algorithm that generates Functional-Class Fingerprints (FCFP), Enhanced Connectivity FingerPrints (ECFP) (Rogers and Hahn 2010) or manually annotated fingerprints such as Molecular ACCess System (MACCS) (Waldrop 1979), Klekota-Roth (KR) (Klekota and F. P. Roth 2008) or PubChem fingerprints. Performance differences for MACCS, ECFP and KR have been described before (Riniker and Landrum 2013; Chee and Oh 2013). The advantage of annotated fingerprints is that

ML algorithm	AUC	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Random Forest	**0.896**	0.886	0.257	0.806	0.886	**0.844**	non-active
		0.743	0.114	**0.845**	0.743	**0.791**	active
Ensemble*	0.895	0.883	**0.253**	**0.808**	0.883	**0.844**	non-active
		0.747	0.117	0.841	**0.747**	**0.791**	active
Bayesian Network	0.878	**0.888**	0.296	0.783	**0.888**	0.832	non-active
		0.704	**0.112**	0.84	0.704	0.766	active
Nave Bayes Tree	0.843	0.839	0.285	0.78	0.839	0.809	non-active
		0.715	0.161	0.787	0.715	0.749	active
K-Nearest Neighbour	0.821	0.818	0.354	0.736	0.818	0.775	non-active
		0.646	0.182	0.747	0.646	0.693	active
LogitBoost	0.82	0.804	0.312	0.756	0.804	0.779	non-active
		0.688	0.196	0.745	0.688	0.715	active
LMT	0.818	0.815	0.297	0.767	0.815	0.791	non-active
		0.703	0.185	0.76	0.703	0.73	active
SMO	0.804	0.801	0.387	0.713	0.801	0.754	non-active
		0.613	0.199	0.719	0.613	0.662	active
J48	0.802	0.824	0.319	0.757	0.824	0.789	non-active
		0.681	0.176	0.763	0.681	0.72	active
Neural Net	0.793	0.711	0.283	0.751	0.711	0.73	non-active
		0.717	0.289	0.673	0.717	0.694	active
LibSVM	0.784	0.793	0.423	0.693	0.793	0.739	non-active
		0.577	0.207	0.698	0.577	0.632	active
Nave Bayes	0.766	0.718	0.349	0.712	0.718	0.715	non-active
		0.651	0.282	0.657	0.651	0.654	active

Table 4.1: Performance of ML algorithms in MolClass. Algorithms ordered by AUC, bold fields indicate best scores in each column and classification category.

structural features are distinct and those that gain more weight by the learning algorithm are interpretable by an expert. Iterative computational fingerprints, due to their incremental nature, produce largely overlapping features that can introduce an additional bias while learning. The full spectrum of fingerprints is implemented in MolClass version 1.5 and the user can choose which fingerprints to use. For example, if a user aims for a model that has the best predictive performance, all features should be combined (see Table 4.2). However, if the main focus is on structure activity relationships, the choice should be an annotated feature vector.

4.4 MolClass predicts molecule preference for efflux pump AcrB

The AcrB dataset has been already introduced in Chapter 3.2.1 to highlight the possibilities to build a model for small molecules that are preferentially pumped. The initial model was developed using Principal Component Analysis (PCA) and a NB learner implemented in Pipeline Pilot. To compare Pipeline Pilot to the abilities in MolClass, the AcrB dataset from Chapter 3.2.1 was used to test some of the classifiers available in MolClass. The performance of four algorithms: a RF, a Bayesian Network, a NB implementation using kernel densities for quantitative values and an ensemble learner with six different learning algorithms are shown in Figure 4.1. The results convey that the prediction accuracy is better or comparable to the naive Bayes implementation in Pipeline Pilot using the initial training set. The random forest and ensemble return both a perfect accuracy raising the concern of overfitting. All MolClass algorithms perform a 5-fold cross validation of the training dataset and the dataset used for the evaluation is highly balanced. The Bayesian network performs slightly better than Pipeline Pilot

Fingerprint	AUC	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
JUMBO	**0.916**	0.907	**0.221**	**0.831**	0.907	**0.868**	non-active
		0.779	0.093	0.875	**0.779**	**0.824**	active
ALL	**0.916**	0.91	0.231	0.826	0.91	0.866	non-active
		0.769	0.09	**0.876**	0.769	0.819	active
MCAT	0.896	0.886	0.257	0.806	0.886	0.844	non-active
		0.743	0.114	0.845	0.743	0.791	active
CDK	0.893	0.879	0.261	0.802	0.879	0.839	non-active
		0.739	0.121	0.836	0.739	0.784	active
Klekota-Roth	0.794	**0.967**	0.646	0.782	**0.967**	0.865	non-active
		0.354	**0.033**	0.815	0.354	0.493	active
EXT	0.778	0.899	0.479	0.819	0.899	0.857	non-active
		0.521	0.101	0.682	0.521	0.59	active
EXTGO	0.774	0.893	0.49	0.814	0.893	0.852	non-active
		0.51	0.107	0.664	0.51	0.577	active
PubChem	0.762	0.927	0.573	0.796	0.927	0.856	non-active
		0.427	0.073	0.709	0.427	0.533	active
MACCS	0.739	0.939	0.636	0.78	0.939	0.852	non-active
		0.364	0.061	0.712	0.364	0.482	active

Table 4.2: Performance of different structural fingerprints for classification. Fingerprints are ordered by AUC, bold fields indicate best scores in each column and classification category. Choice of fingerprints can be used to tailor error rates and sensitivity for minority or majority classification categories.

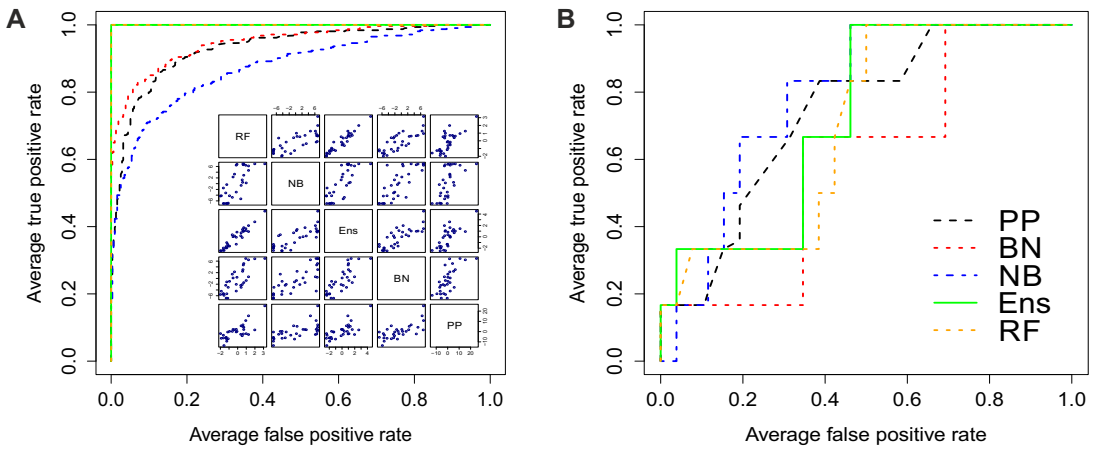


Figure 4.1: MolClass ML models predict molecule transport for AcrB. **A** AUC results from the NB implementation in Pipeline Pilot (PP) and four ML implementations used in MolClass. From MolClass the Bayesian Network (BN), NB, RF, and the ensemble (Ens) learner were tested on the training dataset using 5 fold cross validation. Inlet: Correlation of likelihoods for 32 molecules from the validation dataset for **B** AUC results from an experimental validation of the ML learners to a random set of 32 molecules from a novel vendor library. The training set AUC results in A are 0.93 (PP), 0.95 (BN), 0.78 (NB) and 1 (Ens, RF). The validation AUC results are 0.76 (PP), 0.60 (BN), 0.79 (NB), 0.72 (Ens) and 0.70 (RF).

whereas the NB implementation in MolClass performs slightly worse. As shown in the inset, the performances of all implementations correlate well on a small independent validation dataset indicating differences of the class assignment for each algorithm. The NB implementation in MolClass achieves the highest overall accuracy in the validation dataset as shown in Figure 4.1B. It is important to note that the RF and ensemble implementations show a lower overall accuracy but the molecules with the highest and lowest likelihood values are correctly classified as not pumped and pumped suggesting a weakness in the characterisation of molecules in the midrange. This artefact is due to the nature of the training data as midrange inconclusive molecules were excluded from the training data. Alternatively a regression model for the whole dataset using the quantitative fold suppression values could be a promising alternative versus a binary qualitative model.

4.5 Further development

MolClass supports the analysis of screening data with a broad variety of parametric descriptors and ML algorithms. Some settings are currently restricted, for example the type of cross validation (currently set to 5-fold), the data balancing (Synthetic Minority Over-sampling TEchnique (SMOTE) (Chawla et al. 2002) and subsampling) and correlation-based feature selection (M. A. Hall and Holmes 2003). Important features that I would propose to add to MolClass in the future are score distributions of inter-class and intra-class probabilities for different targets (Keiser, B. L. Roth, et al. 2007), and a unique probability score for each molecule across all models, in addition to an average score over all datasets (Yera, Cleves, and Jain 2011). In its current form, MolClass nevertheless serves as a versatile open source tool to analyse screening data, as demonstrated by several

publications (Burns, Wallace, et al. 2010; Diamandis et al. 2007; Ericson et al. 2008; Hillenmeyer et al. 2008; Ishizaki et al. 2010; Spitzer, Griffiths, et al. 2011).

Chapter 5

Concluding remarks

This thesis is a critical review of my contribution to the development of data analysis procedures to find experimental markers in a wide range of biological studies to accept or reject scientific hypotheses. This work covers the modelling of data, the applicability of existing algorithms and the tailoring of algorithms to specific research questions. The challenge for this thesis was to combine elements of novelty between the biological, statistics and computer sciences. Given the opportunity to be involved in the early stages of collaborative scientific endeavours I believe I was able to provide insights and ideas, which lead to improved experimental designs, better signal to noise ratios and novel biological discoveries.

Throughout this research, the most daunting part was the broad and manifold literature presenting me, the bioinformatician with an overwhelming range of algorithms and their variations to support a biological hypothesis. In retrospect, the work in this thesis is focused on diligent data gathering to support models that can be tested through procedures that increase the data to noise ratio, the control for spatial biases regarding the experimental readouts, the selection of suitable parameters for further analysis or the selection of task-optimised small molecules.

A selected list of methods throughout the research were Principal Component Analysis (PCA) and Bayesian approaches, in particular Naive Bayes (NB). The targeted identification of the sources of bias is illustrated by the different examples across the chapters. Choices on the applied methods were made after exploring the data distributions and appropriate statistical methods. Extensive efforts went into testing and validation of applicable methods and different parameter ranges, in particular to optimise the machine learning approaches.

5.1 Complexity of biological systems

The discrepancies that are frequently observed between different studies underscore the complexity of the underlying biology and the many parameters that can influence screen outputs. The choice of fluorescent markers, cell lines or reagents can drastically influence the outcome of an experiment even if the experimental set-up is founded on the same hypothesis and focused on the same phenotype. Since the reported statistics are usually sound, another factor contributing to discrepancies may be observation bias (Nuzzo 2015). In the review of large datasets, even if results are random, cognitive self-deception can lead researchers to seek out a reasonable result. Some of the intrinsic positive controls, gene functions that have been thoroughly studied, can also provide misplaced confidence in the presumed signal. Extensive mechanism-based follow up experiments are therefore essential to validate any chosen screen hits.

5.2 New statistics for big data

The first decade of the 21st century increasingly requires the processing of large, heterogeneous data sets. Scientists today are not only awash in a flood of data, but

also face a large number of questions concerning any given dataset, a scenario that was not envisioned by the methods developed at the beginning of the last century. Many classical statistical methods assume univariate consistent distributions and small sample sizes. Throughout a scientific workflow a hypothesis was formed: data collected, cleaned, structured and reviewed before the hypothesis was accepted or rejected on statistical grounds. Today's data requires multiple hypothesis tests or even a hypothesis space on an undefined number of data subgroups. As data set size and complexity continues to grow, most successful methods are either non-parametric or allow complex non-linear relationships within the data (Domingos 2012; Fernández-Delgado, Cernadas, and Barro 2014). The current scientific literature promotes an increasing number of complex statistical and machine learning approaches.

5.2.1 Variability in experimental data - accounting for uncertainty

The assessment of variability has important implications since variability in measurements propagate through calculations and inferences. The ultimate goal is to make the reliabilities of its conclusions apparent to the wider audience and leverage global output of the science community. The recent rapid development of high-throughput DNA sequencing technologies has dramatically increased the number of measurements made at the genetic level. These data come from many different DNA-sequencing technologies, each with their own platform-specific errors and biases, which vary widely (O′Rawe, Ferson, and Lyon 2015; Hwang et al. 2015). Similar conclusions can be drawn from the use of imaging platforms, experimental protocols, genetic and fluorescent markers used in high content imaging. Whereas for Next-Generation Sequencing (NGS) several statistical studies have tried to assess error rates for basic work flows between different

technology platforms, such comparisons are outside a feasible scope for a wide range of imaging methods used by experimental biologists. Ideally, the existence of general schemes to project uncertainties on different levels of influencing factors would be beneficial to assess the certainty of the conclusions drawn about biochemical or genetic regulators or more generally, the biological questions. With the success of DNA sequencing applications the trends is clearly towards assessment of uncertainty quantification, to describe sources of error, and propose methods that can be used for accounting and propagating these errors and their uncertainties through subsequent calculations (O′Rawe, Ferson, and Lyon 2015). The Google Baseline and 1000 genome project will be efficient drivers of new statistical models to come. These developments could have a positive effect on other large scale technologies in biology, such as high content imaging, metabolomics and proteomics.

5.2.2 Variability between experimental studies - a call for meta analysis

To increase the confidence in findings from fundamental biological research, it would be beneficial to implement a set of reporting criteria to increase the ability to perform meta-analysis. The application of meta analysis is very common in medical and public health research to verify and standardise treatment procedures and patient care. It is used for medical imaging to define general markers for tumour detection in different tissues. No such studies can be found for localisation patterns of fluorescent markers in biological literature. Given its different nature, Genome-Wide Association Study (GWAS), providing comparable data with an unprecedented number of promising signals of association between genomic variants and human traits have quickly adopted the potential of meta-analysis. The steps required to validate, augment and refine such signals to identify

underlying causal variants for well-defined phenotypes benefit from the collation of different studies. General advantages are the confirmation of association signals through replication. The ability to test whether a signal can be generalised across different populations. The identification of the most informative markers and finding multiple independent markers under each signal. Further, it improves the documentation of biological functions, appreciating the exact phenotypes involved in the association and importantly the detection of potential pleiotropy. The obvious risk of statistical heterogeneity can affect the ability to detect such associations and, when detected, it poses questions about the reasons for its existence. However, one should acknowledge that unless many data sets are combined, there will be a large degree of uncertainty about the amount of estimated between-study heterogeneity (Ioannidis, Thomas, and Daly 2009).

5.2.3 Emphasis on multi-disciplinary scientific reporting

In my opinion, the term big data encapsulates the dilemma that analytical science is currently facing. There is no question that the probabilistic/statistical point of view will continue to dominate science and engineering disciplines. Predictive models and statistical approaches have been used for decades since the advent of modern computers (Metropolis and Ulam 2012) and have become increasingly relevant since most problems with deterministic solutions have been solved. The current popularity of big data builds on the unimaginable large data repositories that now exist, coupled with the demand to make sense of this data and retrieve its intrinsic information. Statistics forms the essence of scientific research and hypothesis testing, but sophisticated analysis procedures are not well established and basic statistical tests and measures, such as the t-test, Z-score, mean and simple counts are commonly used used for quantification. Within large-scale datasets, there is a risk that hits may be selected on the

basis of prior knowledge or experience, resulting in a just-so story if confirmed in any follow up experiment (Nuzzo 2015). The advent of staggeringly large multivariate data sets, brings the danger that these data sets may harbour only weak signals, as illustrated in a meta analysis of gene association studies (Ioannidis, Ntzani, et al. 2001). Statistical methods have barely caught up with such data complexity. The limits of intuition, when looking at dozens or hundreds of variables, makes logical inference impossible. It is important to recognise that the same limitations apply to computational approaches. For some algorithms, generalisation becomes harder with increasing number of dimensions (Bellman 1961). How then can we improve scientific methods to overcome the shortages in data reproducibility and/or stochasticity that are intrinsic to many complex biological responses? To tackle future scientific dataset and research questions, multi-disciplinary collaborative efforts between statistics, machine learning and biological communities are most certainly needed. Consortia can provide a framework to make design choices upfront and harmonise phenotypes and analysis methods. Further, larger coordinated efforts are more likely to wield resources that guarantee that acquired data is robust, representative and transparent. The trend towards multi-disciplinary teams is reflected by the fact that the average number of authors on scientific papers has more than doubled in the past 20 years. Finally, in closing, I would like to emphasise the need for further evolution in the practice of data sharing, analysis and transparency (Peng 2011). The collection of new data, whether it be replication studies, small scale studies, consortia assemblies or meta-analyses, has led to an emerging paradigm of conglomerate analyses (The 1000 Genomes Project Consortium 2010; Xiong et al. 2015). If this approach can be extended to cross current boundaries between GWAS, High Content Screen (HCS), proteomics, metabolomics and drug screen data, systems biology will reach a new qualitative level and offer the hope of truly predictive biological models.

Literature References

Abbas, Syed S, Tjeerd MH Dijkstra, and Tom Heskes (2014). "A comparative study of cell classifiers for image-based high-throughput screening". In: *BMC Bioinformatics* 15.1, p. 342.

Afzal, Avid M et al. (2015). "A multi-label approach to target prediction taking ligand promiscuity into account." In: *Journal of Cheminformatics* 7.1, p. 24.

Apte, Suneel S and William C Parks (2015). "Metalloproteinases: A parade of functions in matrix biology and an outlook for the future". In: *Matrix Biology* 44-46, pp. 1–6.

Baell, Jonathan B and Georgina A Holloway (2010). "New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays". In: *J. Med. Chem.* 53.7, pp. 2719–2740.

Bakal, Chris et al. (2007). "Quantitative morphological signatures define local signaling networks regulating cell morphology." In: *Science* 316.5832, pp. 1753–1756.

Balfer, Jenny et al. (2014). "Modeling of compound profiling experiments using support vector machines." In: *Chem Biol Drug Des* 84.1, pp. 75–85.

Bartlett, M S (1941). "The Statistical Significance of Canonical Correlations". In: *Biometrika* 32.1, pp. 29–37.

Beard, Philippa M et al. (2014). "A loss of function analysis of host factors influencing Vaccinia virus replication by RNA interference." In: *PLoS ONE* 9.6, e98431.

Belenky, Peter, Diogo Camacho, and James J Collins (2013). "Fungicidal Drugs Induce a Common Oxidative-Damage Cellular Death Pathway". In: *Cell Reports* 3.2, pp. 350–358.

Bellman, Richard (1961). "On the reduction of dimensionality for classes of dynamic programming processes". In: *Journal of Mathematical Analysis and Applications* 3.2, pp. 358–360.

Bemis, G W and M A Murcko (1996). "The properties of known drugs .1. Molecular frameworks". In: *J. Med. Chem.* 39.15, pp. 2887–2893.

Berger, Axel B et al. (2008). "High-resolution statistical mapping reveals gene territories in live yeast." In: *Nat Meth* 5.12, pp. 1031–1037.

Bernardo, Diego di et al. (2005). "Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks." In: *Nat Biotechnol* 23.3, pp. 377–383.

Besnard, Jérémy, Philip S Jones, et al. (2015). "The Joint European Compound Library: boosting precompetitive research". In: *Drug Discovery Today* 20.2, pp. 181–186.

Besnard, Jérémy, Gian Filippo Ruda, et al. (2012). "Automated design of ligands to polypharmacological profiles." In: *Nature* 492.7428, pp. 215–220.

Bickerton, G Richard et al. (2012). "Quantifying the chemical beauty of drugs". In: *Nature Chemistry* 4.2, pp. 90–98.

Birmingham, Amanda et al. (2009). "Statistical methods for analysis of high-throughput RNA interference screens". In: *Nat Meth* 6.8, pp. 569–575.

Boland, Michael V and Robert F Murphy (2001). "A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells". In: *Bioinformatics* 17.12, pp. 1213–1223.

Borisy, A A et al. (2003). "Systematic discovery of multicomponent therapeutics". In: *Proceedings of the National Academy of Sciences* 100.13, pp. 7977–7982.

Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer (2011). "D3: Data-Driven Documents". In: *Visualization and Computer Graphics, IEEE Transactions on* 17.12, pp. 2301–2309.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32.

Breinig, Marco et al. (2015). "A chemical–genetic interaction map of small molecules using high-throughput imaging in cancer cells". In: *Molecular Systems Biology* 11.12, pp. 846–846.

Brideau, Christine et al. (2003). "Improved Statistical Methods for Hit Selection in High-Throughput Screening". In: *Journal of Biomolecular Screening* 8.6, pp. 634–647.

Bro, Rasmus and Age K Smilde (2014). "Principal component analysis". In: *Analytical Methods* 6.9, pp. 2812–2831.

Brown, Eric D and Gerard D Wright (2016). "Antibacterial drug discovery in the resistance era". In: *Nature* 529.7586, pp. 336–343.

Buchser, W et al. (2014). "Assay development guidelines for image-based high content screening, high content analysis and high content imaging". In: *Assay Guidance Manual*.

Burg, J P (1967). "Maximum entropy spectral analysis." In: *37th Annual International Meeting*.

Burns, Andrew R, Trevor C Y Kwok, et al. (2006). "High-throughput screening of small molecules for bioactivity and target identification in Caenorhabditis elegans". In: *Nat Protoc* 1.4, pp. 1906–1914.

Burns, Andrew R, Genna M Luciani, et al. (2015). "Caenorhabditis elegans is a useful model for anthelmintic discovery." In: *Nature Communications* 6, p. 7485.

Burns, Andrew R, Iain M Wallace, et al. (2010). "A predictive model for drug bioaccumulation and bioactivity in Caenorhabditis elegans." In: *Nat Chem Biol* 6.7, pp. 549–557.

Campillos, Monica et al. (2008). "Drug target identification using side-effect similarity." In: *Science* 321.5886, pp. 263–266.

Carpenter, Anne E et al. (2006). "CellProfiler: image analysis software for identifying and quantifying cell phenotypes." In: *Genome Biol* 7.10, R100.

Castonguay, Emilie et al. (2015). "Panspecies small-molecule disruptors of heterochromatin-mediated transcriptional gene silencing." In: *Molecular and Cellular Biology* 35.4, pp. 662–674.

Caussinus, Emmanuel and Cayetano Gonzalez (2005). "Induction of tumor growth by altered stem-cell asymmetric division in Drosophila melanogaster". In: *Nat Genet* 37.10, pp. 1125–1129.

Chalfie, M et al. (1994). "Green fluorescent protein as a marker for gene expression." In: *Science* 263.5148, pp. 802–805.

Chatr-aryamontri, Andrew et al. (2015). "The BioGRID interaction database: 2015 update." In: *Nucleic Acids Research* 43.Database issue, pp. D470–8.

Chawla, N V et al. (2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research*, pp. 321–357.

Chee, Hyun Keun and S June Oh (2013). "Molecular vibration-activity relationship in the agonism of adenosine receptors." In: *Genomics Inform* 11.4, pp. 282–288.

Chen, Bin, David Wild, and Rajarshi Guha (2009). "PubChem as a source of polypharmacology." In: *J. Chem. Inf. Model.* 49.9, pp. 2044–2055.

Chen, Yu and Brian K Shoichet (2009). "Molecular docking and ligand specificity in fragment-based inhibitor discovery." In: *Nature Publishing Group* 5.5, pp. 358–364.

Cheng, Feixiong, Weihua Li, et al. (2012). "admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties." In: *J. Chem. Inf. Model.* 52.11, pp. 3099–3105.

Cheng, Feixiong, Chuang Liu, et al. (2012). "Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference". In: *PLoS Comput Biol* 8.5, e1002503.

Cheng, Feixiong and Zhongming Zhao (2014). "Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties." In: *J Am Med Inform Assoc* 21.e2, e278–86.

Chenoweth, M B (1956). "Chelation as a mechanism of pharmacological action." In: *Pharmacol Rev* 8.1, pp. 57–87.

Chong, Curtis R and David J Sullivan (2007). "New uses for old drugs". In: *Nature* 448.7154, pp. 645–646.

Clark, Alex M, Krishna Dole, et al. (2015). "Open Source Bayesian Models. 1. Application to ADME/Tox and Drug Discovery Datasets". In: *J. Chem. Inf. Model.* 55.6, pp. 1231–1245.

Clark, Alex M and Sean Ekins (2015). "Open Source Bayesian Models. 2. Mining a "Big Dataset" To Create and Validate Models with ChEMBL." In: *J. Chem. Inf. Model.* 55.6, pp. 1246–1260.

Collinet, Claudio et al. (2010). "Systems survey of endocytosis by multiparametric image analysis". In: *Nature* 464.7286, pp. 243–249.

Conrad, Christian et al. (2004). "Automatic Identification of Subcellular Phenotypes on Human Cell Arrays". In: *Genome Research* 14.6, pp. 1130–1136.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine Learning* 20.3, pp. 273–297.

Cortes-Ciriano, Isidro, Andreas Bender, and Thérèse E Malliavin (2015). "Comparing the Influence of Simulated Experimental Errors on 12 Machine Learning Algorithms in Bioactivity Modeling Using 12 Diverse Data Sets". In: *J. Chem. Inf. Model.* 55.7, pp. 1413–1425.

Cox, D R (1958). "The Regression Analysis of Binary Sequences". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 20.2, pp. 215–242.

Cully, Megan J and Sally J Leevers (2006). "RNA interference pinpoints regulators of cell size and the cell cycle". In: *Genome Biol* 7.5, pp. 1–3.

Dar, Arvin C et al. (2012). "Chemical genetic discovery of targets and anti-targets for cancer polypharmacology." In: *Nature* 486.7401, pp. 80–84.

Degtyarenko, Kirill et al. (2008). "ChEBI: a database and ontology for chemical entities of biological interest." In: *Nucleic Acids Research* 36.Database issue, pp. D344–50.

Dennis Jr, G et al. (2003). "DAVID: database for annotation, visualization, and integrated discovery". In: *Genome*.

Diamandis, Phedias et al. (2007). "Chemical genetics reveals a complex functional ground state of neural stem cells". In: 3.5, pp. 268–273.

Dittmar, John C et al. (2013). "Physical and genetic-interaction density reveals functional organization and informs significance cutoffs in genome-wide screens". In: *Proceedings of the National Academy of Sciences* 110.18, pp. 7389–7394.

Doil, Carsten et al. (2009). "RNF168 binds and amplifies ubiquitin conjugates on damaged chromosomes to allow accumulation of repair proteins." In: *Cell* 136.3, pp. 435–446.

Domingos, Pedro (2012). "A few useful things to know about machine learning". In: *Commun. ACM* 55.10, pp. 78–87.

Dragiev, Plamen, Robert Nadon, and Vladimir Makarenkov (2011). "Systematic error detection in experimental high-throughput screening." In: 12.1, p. 25.

— (2012). "Two effective methods for correcting experimental high-throughput screening data." In: *Bioinformatics* 28.13, pp. 1775–1782.

Duran-Frigola, Miquel, David Rossell, and Patrick Aloy (2014). "A chemo-centric view of human health and disease." In: *Nature Communications* 5, p. 5676.

Ejim, Linda et al. (2011). "Combinations of antibiotics and nonantibiotic drugs enhance antimicrobial efficacy". In: *Nat Chem Biol* 7.6, pp. 348–350.

Eliceiri, Kevin W et al. (2012). "Biological imaging software tools". In: *Nat Meth* 9.7, pp. 697–710.

Elkington, P T G, C M O'Kane, and J S Friedland (2005). "The paradox of matrix metalloproteinases in infectious disease". In: *Clinical & Experimental Immunology* 142.1, pp. 12–20.

Eom, Young-Woo et al. (2005). "Two distinct modes of cell death induced by doxorubicin: apoptosis and cell death through mitotic catastrophe accompanied by senescence-like phenotype." In: *Oncogene* 24.30, pp. 4765–4777.

Ericson, Elke et al. (2008). "Off-target effects of psychoactive drugs revealed by genome-wide assays in yeast." In: *PLoS Genet.* 4.8, e1000151.

Ermakov, Alexander et al. (2012). "A role for intracellular calcium downstream of G-protein signaling in undifferentiated human embryonic stem cell culture." In: *Stem Cell Res* 9.3, pp. 171–184.

Ernst, Robert et al. (2010). "Multidrug efflux pumps: Substrate selection in ATP-binding cassette multidrug efflux pumps – first come, first served?" In: *FEBS Journal* 277.3, pp. 540–549.

Evans, B E et al. (2002). "Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists". In: *J. Med. Chem.* 31.12, pp. 2235–2246.

Fábián, Ákos, György Vereb, and János Szöllősi (2013). "The hitchhikers guide to cancer stem cell theory: markers, pathways and therapy." In: *Cytometry Part A* 83.1, pp. 62–71.

Farkas, D L et al. (1993). "Multimode light microscopy and the dynamics of molecules, cells, and tissues." In: *Annu. Rev. Physiol.* 55.1, pp. 785–817.

Farmer, Hannah et al. (2005). "Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy". In: *Nature* 434.7035, pp. 917–921.

Fechner, Nikolas et al. (2013). "ChEMBLSpace–a graphical explorer of the chemogenomic space covered by the ChEMBL database." In: *Bioinformatics* 29.4, pp. 523–524.

Fedorov, Yuriy et al. (2006). "Off-target effects by siRNA can induce toxic phenotype." In: *RNA* 12.7, pp. 1188–1196.

Feng, Yan et al. (2009). "Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds." In: *Nat Rev Drug Discov* 8.7, pp. 567–578.

Fernández-Delgado, M, E Cernadas, and S Barro (2014). "Do we need hundreds of classifiers to solve real world classification problems?" In: *The Journal of Machine Learning Research* 15, pp. 3133–3181.

Fire, Andrew et al. (1998). "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans". In: *Nature* 391.6669, pp. 806–811.

Fischer, Bernd et al. (2015). "A map of directional genetic interactions in a metazoan cell." In: *Elife* 4, e05464.

Fix, Evelyn and J. L. Hodges Jr. (1989). "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties". In: *International Statistical Review / Revue Internationale de Statistique* 57.3, pp. 238–247.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2000). "Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors)". In: *The Annals of Statistics* 28.2, pp. 337–407.

Friedman, Nir, Dan Geiger, and Moises Goldszmidt (1997). "Bayesian Network Classifiers". In: *Machine Learning* 29.2-3, pp. 131–163.

Gabriel, K R (1971). "The biplot graphic display of matrices with application to principal component analysis". In: *Biometrika* 58.3, pp. 453–467.

Gaulton, Anna et al. (2012). "ChEMBL: a large-scale bioactivity database for drug discovery." In: *Nucleic Acids Research* 40.Database issue, pp. D1100–7.

Gesztelyi, Rudolf et al. (2012). "The Hill equation and the origin of quantitative pharmacology". In: *Arch. Hist. Exact Sci.* 66.4, pp. 427–438.

Giaever, G et al. (2004). "Chemogenomic profiling: Identifying the functional interactions of small molecules in yeast". In: *Proceedings of the National Academy of Sciences* 101.3, pp. 793–798.

Glick, Meir et al. (2004). "Enrichment of Extremely Noisy High-Throughput Screening Data Using a Naïve Bayes Classifier". In: *Journal of Biomolecular Screening* 9.1, pp. 32–36.

Gordon, Eric M et al. (2002). "Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions". In: *J. Med. Chem.* 37.10, pp. 1385–1401.

Haggarty, Stephen J (2005). "The principle of complementarity: chemical versus biological space". In: *Current Opinion in Chemical Biology* 9.3, pp. 296–303.

Hajduk, P J et al. (2000). "Privileged molecules for protein binding identified from NMR-based screening." In: *J. Med. Chem.* 43.18, pp. 3443–3447.

Hall, Mark A and Geoffrey Holmes (2003). "Benchmarking attribute selection techniques for discrete class data mining". In: *Knowledge and Data Engineering, IEEE Transactions on* 15.6, pp. 1437–1447.

Hammett, Louis P (1935). "Some Relations between Reaction Rates and Equilibrium Constants." In: *Chem. Rev.* 17.1, pp. 125–136.

Hanahan, Douglas (2014). "Rethinking the war on cancer." In: *Lancet* 383.9916, pp. 558–563.

Hansch, C et al. (1962). "Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients". In: *Nature* 194.4824, pp. 178–180.

Haralick, Robert M (1979). "Statistical and structural approaches to texture". In: *Proceedings of the IEEE* 67.5, pp. 786–804.

Hassan, Moises et al. (2006). "Cheminformatics analysis and learning in a data pipelining environment". In: *Mol. Divers.* 10.3, pp. 283–299.

Heikamp, Kathrin and Jürgen Bajorath (2011). "Large-scale similarity search profiling of ChEMBL compound data sets." In: *J. Chem. Inf. Model.* 51.8, pp. 1831–1839.

Heim, R and R Y Tsien (1996). "Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer." In: *Curr. Biol.* 6.2, pp. 178–182.

Held, Michael et al. (2010). "CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging". In: *Nat Meth* 7.9, pp. 747–754.

Herzler, Matthias, Sieglinde Herre, and Fritz Pragst (2003). "Selectivity of substance identification by HPLC-DAD in toxicological analysis using a UV spectra library of 2682 compounds." In: *J Anal Toxicol* 27.4, pp. 233–242.

Hillenmeyer, Maureen E et al. (2008). "The chemical genomic portrait of yeast: uncovering a phenotype for all genes." In: *Science* 320.5874, pp. 362–365.

Hobbs, Errett C et al. (2012). "Conserved small protein associates with the multidrug efflux pump AcrB and differentially affects antibiotic resistance." In: *Proc. Natl. Acad. Sci. U.S.A.* 109.41, pp. 16696–16701.

Hopkins, Andrew L (2008). "Network pharmacology: the next paradigm in drug discovery." In: *Nature Publishing Group* 4.11, pp. 682–690.

— (2009). "Drug discovery: Predicting promiscuity." In: *Nature* 462.7270, pp. 167–168.

Hopkins, Andrew L and G Richard Bickerton (2010). "Drug discovery: Know your chemical space". In: *Nat Chem Biol* 6.7, pp. 482–483.

Horn, Thomas et al. (2011). "Mapping of signaling networks through synthetic genetic interaction analysis by RNAi." In: *Nat Meth* 8.4, pp. 341–346.

Horton, Douglas A, Gregory T Bourne, and Mark L Smythe (2003). "The combinatorial synthesis of bicyclic privileged structures or privileged substructures." In: *Chem. Rev.* 103.3, pp. 893–930.

Hotelling, H (1933). "Analysis of a complex of statistical variables into principal components." In: *Journal of Educational Psychology* 24.6, pp. 417–441.

Hu, Yanhua and Robert F Murphy (2004). "Automated interpretation of subcellular patterns from immunofluorescence microscopy". In: *Journal of Immunological Methods* 290.1-2, pp. 93–105.

Huang, R et al. (2011). "The NCGC Pharmaceutical Collection: A Comprehensive Resource of Clinically Approved Drugs Enabling Repurposing and Chemical Genomics". In: *Science Translational Medicine* 3.80, 80ps16–80ps16.

Hummon, Amanda B et al. (2012). "Systems-wide RNAi analysis of CASP8AP2/FLASH shows transcriptional deregulation of the replication-dependent histone genes and extensive effects on the transcriptome of colorectal cancer cells." In: *Mol. Cancer* 11.1, p. 1.

Hwang, Sohyun et al. (2015). "Systematic comparison of variant calling pipelines using gold standard personal exome variants". In: *Sci. Rep.* 5, p. 17875.

Ihaka, Ross and R Gentleman (1996). "R: A Language for Data Analysis and Graphics". In: *Journal of Computational and Graphical Statistics* 5.3, pp. 299–314.

Ihaka, Ross and Robert Gentleman (2012). "R: A Language for Data Analysis and Graphics". In: *Journal of Computational and Graphical Statistics* 5.3, pp. 299–314.

Ioannidis, John P A, Evangelia E Ntzani, et al. (2001). "Replication validity of genetic association studies". In: *Nat Genet* 29.3, pp. 306–309.

Ioannidis, John P A, Gilles Thomas, and Mark J Daly (2009). "Validating, augmenting and refining genome-wide association signals". In: *Nat Rev Genet* 10.5, pp. 318–329.

Irwin, John J et al. (2012). "ZINC: A Free Tool to Discover Chemistry for Biology". In: *J. Chem. Inf. Model.* 52.7, pp. 1757–1768.

Ishizaki, Hironori et al. (2010). "Combined zebrafish-yeast chemical-genetic screens reveal gene-copper-nutrition interactions that modulate melanocyte pigmentation." In: *Dis Model Mech* 3.9-10, pp. 639–651.

Jackson, Aimee L and Peter S Linsley (2010). "Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application". In: *Nat Rev Drug Discov* 9.1, pp. 57–67.

Jia, Jia et al. (2009). "Mechanisms of drug combinations: interaction and network perspectives." In: *Nat Rev Drug Discov* 8.2, pp. 111–128.

Jones, Peter A and Stephen B Baylin (2007). "The Epigenomics of Cancer". In: *Cell* 128.4, pp. 683–692.

Jones, Thouis R, Anne E Carpenter, et al. (2009). "Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning." In: *Proc. Natl. Acad. Sci. U.S.A.* 106.6, pp. 1826–1831.

Jones, Thouis R, In Han Kang, et al. (2008). "CellProfiler Analyst: data exploration and analysis software for complex image-based screens." In: *BMC Bioinformatics* 9.1, p. 482.

Keiser, Michael J, Bryan L Roth, et al. (2007). "Relating protein pharmacology by ligand chemistry." In: *Nat Biotechnol* 25.2, pp. 197–206.

Keiser, Michael J, Vincent Setola, et al. (2009). "Predicting new molecular targets for known drugs." In: *Nature* 462.7270, pp. 175–181.

Kim, Tae-Kyun, Josef Kittler, and Roberto Cipolla (2007). "Discriminative learning and recognition of image set classes using canonical correlations." In: *IEEE Trans Pattern Anal Mach Intell* 29.6, pp. 1005–1018.

Kitano, Hiroaki (2007). "A robustness-based approach to systems-oriented drug design." In: *Nat Rev Drug Discov* 6.3, pp. 202–210.

Kittler, Ralf et al. (2007). "Genome-wide resources of endoribonuclease-prepared short interfering RNAs for specific loss-of-function studies". In: *Nat Meth* 4.4, pp. 337–344.

Kiyohara, Shunsuke et al. (2011). "Tryptophan auxotroph mutants suppress the superroot2 phenotypes, modulating IAA biosynthesis in arabidopsis." In: *Plant Signal Behav* 6.9, pp. 1351–1355.

Klekota, Justin and Frederick P Roth (2008). "Chemical substructures that enrich for biological activity." In: *Bioinformatics* 24.21, pp. 2518–2525.

Koh, Judice L Y et al. (2015). "CYCLoPs: A Comprehensive Database Constructed from Automated Analysis of Protein Abundance and Subcellular Localization Patterns in Saccharomyces cerevisiae." In: *G3 (Bethesda)* 5.6, pp. 1223–1232.

Kohonen, Pekka et al. (2013). "The ToxBank Data Warehouse: Supporting the Replacement of In Vivo Repeated Dose Systemic Toxicity Testing". In: *Mol. Inf.* 32.1, pp. 47–63.

Kolas, Nadine K et al. (2007). "Orchestration of the DNA-Damage Response by the RNF8 Ubiquitin Ligase". In: *Science* 318.5856, pp. 1637–1640.

Kolmogoroff, A (1941). "Confidence limits for an unknown distribution function". In: *The Annals of Mathematical Statistics*.

Koutsoukas, Alexios et al. (2013). "In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt window." In: *J. Chem. Inf. Model.* 53.8, pp. 1957–1966.

Krier, Mireille, Guillaume Bret, and Didier Rognan (2006). "Assessing the scaffold diversity of screening libraries." In: *J. Chem. Inf. Model.* 46.2, pp. 512–524.

Krogan, Nevan J et al. (2004). "Proteasome Involvement in the Repair of DNA Double-Strand Breaks". In: *Molecular Cell* 16.6, pp. 1027–1034.

Kuhn, Michael et al. (2013). "Systematic identification of proteins that elicit drug side effects". In: *Mol. Syst. Biol.* 9.1, pp. 663–663.

Kuhn, M et al. (2011). "STITCH 3: zooming in on protein-chemical interactions". In: *Nucleic Acids Research* 40.D1, pp. D876–D880.

Kuncheva, Ludmila I and Christopher J Whitaker (2003). "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy". In: *Machine Learning* 51.2, pp. 181–207.

Kutchukian, Peter S et al. (2016). "Chemistry informer libraries: a chemoinformatics enabled approach to evaluate and advance synthetic methods". In: *Chemical Science* 7.4, pp. 2604–2613.

Lamb, J (2006). "The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease". In: *Science* 313.5795, pp. 1929–1935.

Lander, E S et al. (2001). "Initial sequencing and analysis of the human genome." In: *Nature* 409.6822, pp. 860–921.

Landwehr, Niels, Mark Hall, and Eibe Frank (2005). "Logistic Model Trees". In: *Machine Learning* 59.1-2, pp. 161–205.

Lassmann, Michael et al. (2010). "In Vivo Formation of γ-H2AX and 53BP1 DNA Repair Foci in Blood Cells After Radioiodine Therapy of Differentiated Thyroid Cancer". In: *J. Nucl. Med.* 51.8, pp. 1318–1325.

Laufer, Christina et al. (2013). "Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping." In: *Nat Meth* 10.5, pp. 427–431.

Law, Vivian et al. (2014). "DrugBank 4.0: shedding new light on drug metabolism." In: *Nucleic Acids Research* 42.Database issue, pp. D1091–7.

Lee, Anna Y et al. (2014). "Mapping the cellular response to small molecules using chemogenomic fitness signatures." In: *Science* 344.6180, pp. 208–211.

Lehár, Joseph, Andrew S Krueger, et al. (2009). "Synergistic drug combinations tend to improve therapeutically relevant selectivity." In: *Nat Biotechnol* 27.7, pp. 659–666.

Lehár, Joseph, Andrew Krueger, et al. (2008). "High-order combination effects and biological robustness." In: *Molecular Systems Biology* 4, p. 215.

Lehár, Joseph, Grant R Zimmermann, et al. (2007). "Chemical combination effects predict connectivity in biological systems". In: *Mol. Syst. Biol.* 3, p. 80.

Li, Xiaoming et al. (2004). "Multicopy Suppressors for Novel Antibacterial Compounds Reveal Targets and Drug Efflux Susceptibility". In: *Chemistry & Biology* 11.10, pp. 1423–1430.

Liberali, Prisca, Berend Snijder, and Lucas Pelkmans (2014). "A hierarchical map of regulatory genetic interactions in membrane trafficking." In: *Cell* 157.6, pp. 1473–1487.

— (2015). "Single-cell and multivariate approaches in genetic perturbation screens." In: *Nat Rev Genet* 16.1, pp. 18–32.

Lipinski, C A et al. (2001). "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings." In: *Adv. Drug Deliv. Rev.* 46.1-3, pp. 3–26.

Lipinski, Christopher and Andrew Hopkins (2004). "Navigating chemical space for biology and medicine". In: *Nature* 432.7019, pp. 855–861.

Lönnstedt, I and T Speed (2002). "Replicated microarray data". In: *Statistica sinica*.

Lynch, Michael F, John M Barnard, and Stephen M Welford (1981). "Computer storage and retrieval of generic chemical structures in patents. 1. Introduction and general strategy". In: *J. Chem. Inf. Model.* 21.3, pp. 148–150.

Machida, Yuichi J and Anindya Dutta (2007). "The APC/C inhibitor, Emi1, is essential for prevention of rereplication." In: *Genes & Development* 21.2, pp. 184–194.

Mahalanobis, P C (1936). *On the generalized distance in statistics*. Proceedings of the National Institute of Sciences (India).

Makarenkov, Vladimir et al. (2006). "HTS-Corrector: software for the statistical analysis and correction of experimental high-throughput screening data." In: *Bioinformatics* 22.11, pp. 1408–1409.

Malo, Nathalie et al. (2006). "Statistical practice in high-throughput screening data analysis." In: *Nat Biotechnol* 24.2, pp. 167–175.

Maron, M E (1961). "Automatic Indexing: An Experimental Inquiry". In: *Journal of the ACM (JACM)* 8.3, pp. 404–417.

Mason, J S et al. (1999). "New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures." In: *J. Med. Chem.* 42.17, pp. 3251–3264.

Massey Jr, Frank J (2012). "The Kolmogorov-Smirnov Test for Goodness of Fit". In: *Journal of the American Statistical Association* 46.253, pp. 68–78.

Masunov, Artem E and Ivan A Mikhailov (2010). "Theory and computations of two-photon absorbing photochromic chromophores". In: *European Journal of Chemistry* 1.2, pp. 142–161–161.

McDermott, Sean P and Max S Wicha (2010). "Targeting breast cancer stem cells". In: *Molecular Oncology* 4.5, pp. 404–419.

McQuitty, Louis L (1960). "Hierarchical Linkage Analysis for the Isolation of Types". In: *Educational and Psychological Measurement* 20.1, pp. 55–67.

Mercer, Jason et al. (2012). "RNAi Screening Reveals Proteasome- and Cullin3-Dependent Stages in Vaccinia Virus Infection". In: *Cell Reports* 2.4, pp. 1036–1047.

Metropolis, Nicholas and S Ulam (2012). "The Monte Carlo Method". In: *Journal of the American Statistical Association* 44.247, pp. 335–341.

Mi, Huaiyu et al. (2005). "The PANTHER database of protein families, subfamilies, functions and pathways". In: *Nucleic Acids Research* 33.suppl 1, pp. D284–D288.

Millard, Bjorn L et al. (2011). "Adaptive informatics for multifactorial and high-content biological data". In: *Nat Meth* 8.6, pp. 487–492.

Miozzari, G, P Niederberger, and R Hütter (1978). "Tryptophan biosynthesis in Saccharomyces cerevisiae: control of the flux through the pathway." In: *J. Bacteriol.* 134.1, pp. 48–59.

Mizera, Mikołaj et al. (2015). "Prediction of HPLC retention times of tebipenem pivoxyl and its degradation products in solid state by applying adaptive artificial neural network with recursive features elimination". In: *Talanta* 137, pp. 174–181.

Morgan, H L (1965). "The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service." In: *J. Chem. Doc.* 5.2, pp. 107–113.

Mostafavi, S et al. (2008). "GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function". In: *Genome Biol.*

Muller, Keith E (1981). "Relationships between redundancy analysis, canonical correlation, and multivariate regression". In: *Psychometrika* 46.2, pp. 139–142.

Nelander, Sven et al. (2008). "Models from experiments: combinatorial drug perturbations of cancer cells." In: *Molecular Systems Biology* 4, p. 216.

Nidhi et al. (2006). "Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases." In: *J. Chem. Inf. Model.* 46.3, pp. 1124–1133.

Nikaido, H and H I Zgurskaya (2001). "AcrAB and related multidrug efflux pumps of Escherichia coli." In: *J. Mol. Microbiol. Biotechnol.* 3.2, pp. 215–218.

Nocito, Antonio et al. (2008). "Serotonin regulates macrophage-mediated angiogenesis in a mouse model of colon cancer allografts." In: *Cancer Research* 68.13, pp. 5152–5158.

Nuzzo, Regina (2015). "How scientists fool themselves - and how they can stop." In: *Nature* 526.7572, pp. 182–185.

O'Boyle, Noel M et al. (2011). "Open Babel: An open chemical toolbox." In: *Journal of Cheminformatics* 3.1, p. 33.

Odén, A and H Wedel (1975). "Arguments for Fisher's Permutation Test on JSTOR". In: *The Annals of Statistics*.

O'Donnell, Lara et al. (2010). "The MMS22L-TONSL complex mediates recovery from replication stress and homologous recombination." In: *Molecular Cell* 40.4, pp. 619–631.

Ogier, Arnaud and Thierry Dorval (2012). "HCS-Analyzer: open source software for high-content screening data correction and analysis." In: *Bioinformatics* 28.14, pp. 1945–1946.

Olah, M et al. (2008). "WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery". In: *Chemical Biology From Small Molecules to Systems Biology and Drug Design* 1-3, pp. 760–786.

O′Neill, John S et al. (2011). "Circadian rhythms persist without transcription in a eukaryote". In: *Nature* 469.7331, pp. 554–558.

Oprea, Tudor I et al. (2009). "A crowdsourcing evaluation of the NIH chemical probes". In: *Nat Chem Biol* 5.7, pp. 441–447.

O′Rawe, Jason A, Scott Ferson, and Gholson J Lyon (2015). "Accounting for uncertainty in DNA sequencing data". In: *Trends in Genetics* 31.2, pp. 61–66.

Overington, John P, Bissan Al-Lazikani, and Andrew L Hopkins (2006). "How many drug targets are there?" In: *Nat Rev Drug Discov* 5.12, pp. 993–996.

Panier, Stephanie and Daniel Durocher (2009). "Regulatory ubiquitylation in response to DNA double-strand breaks". In: *DNA Repair (Amst.)* 8.4, pp. 436–443.

Paolini, Gaia V et al. (2006). "Global mapping of pharmacological space." In: *Nat Biotechnol* 24.7, pp. 805–815.

Park, P J et al. (2009). "A permutation test for determining significance of clusters with applications to spatial and gene expression data". In: *Computational Statistics & Data Analysis* 53.12, pp. 4290–4300.

Parkinson, H et al. (2007). "ArrayExpress–a public database of microarray experiments and gene expression profiles." In: *Nucleic Acids Research* 35.Database issue, pp. D747–50.

Parsons, Ainslie B, Renee L Brost, et al. (2003). "Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways". In: *Nat Biotechnol* 22.1, pp. 62–69.

Parsons, Ainslie B, Andres Lopez, et al. (2006). "Exploring the Mode-of-Action of Bioactive Compounds by Chemical-Genetic Profiling in Yeast". In: *Cell* 126.3, pp. 611–625.

Patterson, Stephen and Susan Wyllie (2014). "Nitro drugs for the treatment of trypanosomatid diseases: past, present, and future prospects". In: *Trends in Parasitology* 30.6, pp. 289–298.

Pau, Gregoire et al. (2010). "EBImage–an R package for image processing with applications to cellular phenotypes." In: *Bioinformatics* 26.7, pp. 979–981.

Paulsen, Renee D et al. (2009). "A genome-wide siRNA screen reveals diverse cellular processes and pathways that mediate genome stability." In: *Molecular Cell* 35.2, pp. 228–239.

Pelz, Oliver, Moritz Gilsdorf, and Michael Boutros (2010). "web cellHTS2: a web-application for the analysis of high-throughput screening data." In: *BMC Bioinformatics* 11, p. 185.

Peng, Roger D (2011). "Reproducible Research in Computational Science". In: *Science* 334.6060, pp. 1226–1227.

Perkins, Theodore J and Peter S Swain (2009). "Strategies for cellular decision-making". In: *Molecular Systems Biology* 5.1.

Petri Seiler, Kathleen et al. (2011). "Master Data Management: Getting your House in Order". In: 14.9, pp. 749–756.

Petrone, Paula M et al. (2012). "Rethinking molecular similarity: comparing compounds on the basis of biological activity." In: *ACS Chem. Biol.* 7.8, pp. 1399–1409.

Pollard, Steven M et al. (2009). "Glioma Stem Cell Lines Expanded in Adherent Culture Have Tumor-Specific Phenotypes and Are Suitable for Chemical and Genetic Screens". In: *Cell Stem Cell* 4.6, pp. 568–580.

Qiu, Peng et al. (2011). "Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE". In: *Nat Biotechnol* 29.10, 886–U181.

Quackenbush, John (2002). "Microarray data normalization and transformation". In: *Nat Genet* 32.Supp, pp. 496–501.

Quinlan, J R (1986). "Induction of Decision Trees". In: *Machine Learning* 1.1, pp. 81–106.

Raju, T N (2000). "The Nobel chronicles. 1988: James Whyte Black, (b 1924), Gertrude Elion (1918-99), and George H Hitchings (1905-98)." In: *Lancet* Mar 18;355(9208), p. 1022.

Ranganathan, Prathibha, Kelly L Weaver, and Anthony J Capobianco (2011). "Notch signalling in solid tumours: a little bit of everything but not all the time." In: *Nature Reviews Cancer* 11.5, pp. 338–351.

Ray, L C and R A Kirsch (1957). "Finding Chemical Records by Digital Computers." In: *Science* 126.3278, pp. 814–819.

Reed, Russell D and Robert J Marks (1998). *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. MIT Press.

Reymond, Jean-Louis et al. (2010). "Chemical space as a source for new drugs". In: *Med. Chem. Commun.* 1.1, pp. 30–38.

Rieber, Nora et al. (2009). "RNAither, an automated pipeline for the statistical analysis of high-throughput RNAi screens." In: *Bioinformatics* 25.5, pp. 678–679.

Riniker, Sereina and Gregory A Landrum (2013). "Open-source platform to benchmark fingerprints for ligand-based virtual screening." In: *J Cheminf* 5.1, p. 26.

Roemer, Terry T and Charles C Boone (2013). "Systems-level antimicrobial drug and drug synergy discovery." In: *Nat. Chem. Biol.* 9.4, pp. 222–231.

Rogers, David, Robert D Brown, and Mathew Hahn (2005). "Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up." In: 10.7, pp. 682–686.

Rogers, David and Mathew Hahn (2010). "Extended-connectivity fingerprints." In: *J. Chem. Inf. Model.* 50.5, pp. 742–754.

Rubí, Blanca and Pierre Maechler (2013). "Minireview: New Roles for Peripheral Dopamine on Metabolic Control and Tumor Growth: Let′s Seek the Balance". In: *Endocrinology* 151.12, pp. 5570–5581.

Sachlos, Eleftherios et al. (2012). "Identification of Drugs Including a Dopamine Receptor Antagonist that Selectively Target Cancer Stem Cells". In: *Cell* 149.6, pp. 1284–1297.

Sailem, Heba Z, Julia E Sero, and Chris Bakal (2015). "Visualizing cellular imaging data using PhenoPlot." In: *Nature Communications* 6, p. 5825.

Scemama, J (1984). "Dopamine receptors in a human colonic cancer cell line (HT29). Some receptor-related biological effects of dopamine". In: *International Journal of Cancer* 34.5, pp. 675–679.

Schierz, Amanda C (2009). "Virtual screening of bioassay data." In: *Journal of Cheminformatics* 1, p. 21.

Schindelin, Johannes et al. (2012). "Fiji: an open-source platform for biological-image analysis". In: *Nat Meth* 9.7, pp. 676–682.

Schneider, Caroline A, Wayne S Rasband, and Kevin W Eliceiri (2012). "NIH Image to ImageJ: 25 years of image analysis". In: *Nat Meth* 9.7, pp. 671–675.

Seiler, Kathleen Petri et al. (2008). "ChemBank: a small-molecule screening and cheminformatics resource database." In: *Nucleic Acids Research* 36.Database issue, pp. D351–9.

Shannon, Paul et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." In: *Genome Research* 13.11, pp. 2498–2504.

Shepard, Roger N (1980). "Multidimensional Scaling, Tree-Fitting, and Clustering". In: *Science* 210.4468, pp. 390–398.

Sheridan, Robert P and Simon K Kearsley (2002). "Why do we need so many chemical similarity search methods?" In: *Drug Discovery Today* 7.17, pp. 903–911.

Siddiqui, Emad J et al. (2006). "The Role of Serotonin (5-Hydroxytryptamine1A and 1B) Receptors in Prostate Cancer Cell Proliferation". In: *The Journal of Urology* 176.4, pp. 1648–1653.

Singh, Dinesh Kumar et al. (2010). "Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities". In: *Molecular Systems Biology* 6.1.

Singh, S, A E Carpenter, and A Genovesio (2014). "Increasing the Content of High-Content Screening: An Overview". In: *Journal of Biomolecular Screening* 19.5, pp. 640–650.

Siracusa, Michael R et al. (2005). "Estimating dependency and significance for high-dimensional data". In: *(ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* 5, pp. 1085–1088.

Slacka, Michael D et al. (2008). "Characterizing heterogeneous cellular responses to perturbations". In: *Proceedings of the National Academy of Sciences* 105.49, pp. 19306–19311.

Smirnov, N (1948). "Table for Estimating the Goodness of Fit of Empirical Distributions on JSTOR". In: *The Annals of Mathematical Statistics*.

Smyth, Gordon K, Yee Hwa Yang, and Terry Speed (2003). "Statistical Issues in cDNA Microarray Data Analysis". In: *Functional Genomics*. New Jersey: Humana Press, pp. 111–136.

Snijder, Berend, Raphael Sacher, Pauli Rämö, Eva-Maria Damm, et al. (2009). "Population context determines cell-to-cell variability in endocytosis and virus infection." In: *Nature* 461.7263, pp. 520–523.

Snijder, Berend, Raphael Sacher, Pauli Rämö, Prisca Liberali, et al. (2012). "Single-cell analysis of population context advances RNAi screening at multiple levels." In: *Molecular Systems Biology* 8, p. 579.

Sokolove, P G and W N Bushell (1978). "The chi square periodogram: its utility for analysis of circadian rhythms." In: *Journal of Theoretical Biology* 72.1, pp. 131–160.

Soll, Christopher, Jae Hwi Jang, et al. (2010). "Serotonin promotes tumor growth in human hepatocellular cancer." In: *Hepatology* 51.4, pp. 1244–1254.

Soll, Christopher, Marc-Oliver Riener, et al. (2012). "Expression of serotonin receptors in human hepatocellular cancer." In: *Clin Cancer Res* 18.21, pp. 5902–5910.

Sommer, Christoph, Michael Held, et al. (2013). "CellH5: a format for data exchange in high-content screening." In: *Bioinformatics* 29.12, pp. 1580–1582.

Sommer, Christoph, Christoph Straehle, et al. (2011). "Ilastik: Interactive learning and segmentation toolkit". In: *IEEE International Symposium on Biomedical Imaging From Nano to Macro*, pp. 230–233.

Southan, Christopher, Péter Várkonyi, and Sorel Muresan (2009). "Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds." In: *Journal of Cheminformatics* 1.1, pp. 10–10.

Spitzer, Michaela, Emma Griffiths, et al. (2011). "Cross-species discovery of syncretic drug combinations that potentiate the antifungal fluconazole." In: *Mol. Syst. Biol.* 7.1, pp. 499–499.

Spitzer, Michaela, Jan Wildenhain, et al. (2014). "BoxPlotR: a web tool for generation of box plots". In: *Nat Meth* 11.2, pp. 121–122.

Stanstrup, Jan, Steffen Neumann, and Urška Vrhovšek (2015). "PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems." In: *Anal. Chem.* 87.18, pp. 9421–9428.

Statnikov, Alexander, Lily Wang, and Constantin F Aliferis (2008). "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification". In: *BMC Bioinformatics* 9.1, p. 319.

Stepanova, Anna N et al. (2011). "The Arabidopsis YUCCA1 flavin monooxygenase functions in the indole-3-pyruvic acid branch of auxin biosynthesis." In: *Plant Cell* 23.11, pp. 3961–3973.

Stephan, Johannes, Oliver Stegle, and Andreas Beyer (2015). "A random forest approach to capture genetic effects in the presence of population structure." In: *Nature Communications* 6, p. 7432.

Stewart, Grant S, Stephanie Panier, et al. (2009). "The RIDDLE syndrome protein mediates a ubiquitin-dependent signaling cascade at sites of DNA damage." In: *Cell* 136.3, pp. 420–434.

Stewart, Grant S, Tatjana Stankovic, et al. (2007). "RIDDLE immunodeficiency syndrome is linked to defects in 53BP1-mediated DNA damage signaling." In: *Proceedings of the National Academy of Sciences* 104.43, pp. 16910–16915.

Straume, Martin, Susan G Frasier-Cadoret, and Michael L Johnson (2002). "Least-Squares Analysis of Fluorescence Data". In: *Topics in Fluorescence Spectroscopy*. Boston: Springer US, pp. 177–240.

Strauss, Michael J (1979). "The nitroaromatic group in drug design. Pharmacology and toxicology (for nonpharmacologists)". In: *Industrial and Engineering Chemistry Product Research and Development* 18.3, pp. 158–166.

Strebhardt, Klaus and Axel Ullrich (2008). "Paul Ehrlich's magic bullet concept: 100 years of progress". In: *Nature Reviews Cancer* 8.6, pp. 473–480.

Tan, Brenton Thomas et al. (2006). "The cancer stem cell hypothesis: a work in progress." In: *Lab. Invest.* 86.12, pp. 1203–1207.

Tanaka, Masahiro et al. (2005). "An unbiased cell morphology-based screen for new, biologically active small molecules." In: *PLoS Biol* 3.5, e128.

Teodori, E et al. (2002). "The medicinal chemistry of multidrug resistance (MDR) reversing drugs". In: *Il Farmaco* 57.5, pp. 385–415.

The 1000 Genomes Project Consortium (2010). "A map of human genome variation from population-scale sequencing". In: *Nature* 467.7319, pp. 1061–1073.

Torgerson, Warren S (1952). "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4, pp. 401–419.

Visnyei, Koppany et al. (2011). "A molecular screening approach to identify and characterize inhibitors of glioblastoma stem cells." In: *Molecular Cancer Therapeutics* 10.10, pp. 1818–1828.

Vleduts, G E (1963). "Concerning one system of classification and codification of organic reactions". In: *Information Storage and Retrieval* 1.2-3, pp. 117–146.

Wagner, Bridget K and Paul A Clemons (2009). "Connecting synthetic chemistry decisions to cell and genome biology using small-molecule phenotypic profiling". In: *Current Opinion in Chemical Biology* 13.5-6, pp. 539–548.

Waldrop, Mitch (1979). "Company Offers Computer-Assisted Chemistry." In: *Chemical and Engineering News* 57.25, p. 29.

Wallace, Iain M et al. (2011). "Compound Prioritization Methods Increase Rates of Chemical Probe Discovery in Model Organisms". In: *Chemistry & Biology* 18.10, pp. 1273–1283.

Wang, Yanli, Evan Bolton, et al. (2010). "An overview of the PubChem BioAssay resource." In: *Nucleic Acids Research* 38.Database issue, pp. D255–66.

Wang, Yanli, Jewen Xiao, et al. (2012). "PubChem's BioAssay Database." In: *Nucleic Acids Research* 40.Database issue, pp. D400–12.

Weininger, David (1988). "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". In: *J. Chem. Inf. Model.* 28.1, pp. 31–36.

Welch, B L (1947). "The Generalization of 'Student's' Problem when Several Different Population Variances are Involved". In: *Biometrika* 34.1/2, p. 28.

Westen, Gerard J P van and John P Overington (2013). "A ligand's-eye view of protein similarity". In: *Nat Meth* 10.2, pp. 116–117.

Wilcoxon, Frank (1945). "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1.6, p. 80.

Wildenhain, Jan and Edmund J Crampin (2006). "Reconstructing gene regulatory networks: from random to scale-free connectivity". In: *IET Sys. Bio.* 153.4, pp. 247–256.

Wildenhain, Jan, Nicholas Fitzgerald, and Mike Tyers (2012). "MolClass: a web portal to interrogate diverse small molecule screen datasets with different computational models." In: *Bioinformatics* 28.16, pp. 2200–2201.

Wildenhain, Jan, Michaela Spitzer, David S Bellows, et al. (2015). "Machine learning uncovers chemical synergies that surmount buffering in the genetic landscape". In: *Cell Systems* 6, pp. 1–13.

Wildenhain, Jan, Michaela Spitzer, Sonam Dolma, et al. (2015). "Prediction of Synergism from Chemical-Genetic Interactions by Machine Learning". In: *Cell Systems* 1.6, pp. 383–395.

Winter, Andrew G, Jan Wildenhain, and Mike Tyers (2011). "BioGRID REST Service, BiogridPlugin2 and BioGRID WebGraph: new tools for access to interaction data at BioGRID." In: *Bioinformatics* 27.7, pp. 1043–1044.

Winzeler, E A et al. (1999). "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis." In: *Science* 285.5429, pp. 901–906.

Wishart, David S (2008). "DrugBank and its relevance to pharmacogenomics." In: *Pharmacogenomics* 9.8, pp. 1155–1162.

Wolfender, Jean-Luc (2009). "HPLC in natural product analysis: the detection issue." In: *Planta Med.* 75.7, pp. 719–734.

Wong, Lai Hong et al. (2013). "A yeast chemical genetic screen identifies inhibitors of human telomerase." In: *Chemistry & Biology* 20.3, pp. 333–340.

Woodward, Wendy A et al. (2007). "WNT/beta-catenin mediates radiation resistance of mouse mammary progenitor cells." In: *Proceedings of the National Academy of Sciences* 104.2, pp. 618–623.

Workman, Paul and Ian Collins (2010). "Probing the Probes: Fitness Factors For Small Molecule Tools". In: *Chemistry & Biology* 17.6, pp. 561–577.

Xia, Xiaoyang et al. (2004). "Classification of Kinase Inhibitors Using a Bayesian Model". In: *J. Med. Chem.* 47.18, pp. 4463–4470.

Xiong, Hui Y et al. (2015). "RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease." In: *Science* 347.6218, pp. 1254806–1254806.

Yabuuchi, Hiroaki et al. (2011). "Analysis of multiple compound-protein interactions reveals novel bioactive molecules." In: *Molecular Systems Biology* 7, p. 472.

Yera, Emmanuel R, Ann E Cleves, and Ajay N Jain (2011). "Chemical structural novelty: on-targets and off-targets." In: *J. Med. Chem.* 54.19, pp. 6771–6785.

Yildirim, Muhammed A et al. (2007). "Drug-target network." In: *Nat Biotechnol* 25.10, pp. 1119–1126.

Yin, Zheng et al. (2013). "A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes." In: *Nat. Cell Biol.* 15.7, pp. 860–871.

Yu, Edward W et al. (2003). "Structural Basis of Multiple Drug-Binding Capacity of the AcrB Multidrug Efflux Pump". In: *Science* 300.5621, pp. 976–980.

Yu, Hua et al. (2012). "A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data." In: *PLoS ONE* 7.5, e37608–e37608.

Zabawa, Thomas P et al. (2016). "Treatment of Gram-negative bacterial infections by potentiation of antibiotics". In: *Current Opinion in Microbiology* 33, pp. 7–12.

Zhang, Ji-Hu, Thomas D Y Chung, and Kevin R Oldenburg (1999). "A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays". In: *Journal of Biomolecular Screening* 4.2, pp. 67–73.

Zhang, Xiaohua Douglas (2007). "A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays". In: *Genomics* 89.4, pp. 552–561.

Zhou, Linna et al. (2012). "ALDH2 mediates 5-nitrofuran activity in multiple species." In: *Chemistry & Biology* 19.7, pp. 883–892.

Zhou, Shu-Feng, Jun-Ping Liu, and Balram Chowbay (2009). "Polymorphism of human cytochrome P450 enzymes and its clinical impact". In: *Drug Metab. Rev.* 41.2, pp. 89–295.

Zhou, Xin et al. (2014). "Development of an in vitro screen for compound bioaccumulation in Haemonchus contortus." In: *J. Parasitol.* 100.6, pp. 848–855.

Appendix A

List of Co-authored Publications

This list showcases my academic output with a brief description of my contribution for each publication.

A.1 Selected Publications

- Nadine K. Kolas, J. Ross Chapman, Shinichiro Nakada, Jarkko Ylanko, Richard Chahwan, Frédéric D. Sweeney, Stephanie Panier, Megan Mendez, **Jan Wildenhain**, Timothy M. Thomson, Laurence Pelletier, Stephen P. Jackson, and Daniel Durocher. Orchestration of the DNA-Damage Response by the RNF8 Ubiquitin Ligase. *Science*, 318(5856):1637–1640, December 2007.

  In collaboration with Nadine Kolas and Jarkko Ylanko I analysed and refined the High Content Screen (HCS) as described in Chapter 2.

- Lara O'Donnell*, Stephanie Panier*, **Jan Wildenhain***, Johnny M Tkach, Abdallah Al-Hakim, Marie-Claude Landry, Cristina Escribano-Diaz, Rachel K. Szilard, Jordan T. F. Young, Meagan Munro, Marella D. Canny, Nadine K. Kolas, Wei Zhang, Shane M. Harding, Jarkko Ylanko, Megan Mendez, Michael Mullin, Thomas Sun, Bianca Habermann, Alessandro Datti, Robert G. Bristow, Anne-Claude Gingras, Michael D. Tyers, Grant W. Brown, and Daniel Durocher. The MMS22L-TONSL complex mediates recovery from replication stress and homologous recombination. *Molecular Cell*, 40(4):619–631, November 2010.

  The list of uncharacterised Open Reading Frame (ORF)s contained C6orf167 which we discovered in this work to be the *human* homolog of Methyl Methanesulfonate Sensitivity 22 (MMS22) from yeast. My direct

contributions to this publication include the analysis of the Fluorescence-Activated Cell Sorting (FACS) profile primary data in Figure 1A, 3A and E. Further, I performed the analysis for the data shown in Figure 4B, C and E. I also provided data and text for the supplementary information and tables.

- Andrew R. Burns, Iain M. Wallace, **Jan Wildenhain**, Mike Tyers, Guri Giaever, Gary D. Bader, Corey Nislow, Sean R. Cutler, and Peter J. Roy. A predictive model for drug bioaccumulation and bioactivity in *Caenorhabditis elegans*. *Nature Chemical Biology*, 6(7):549–557, July 2010.

  Thesis Chapter 3 contains details of the workflows and analysis used in this study and a critical retrospective review. I provided input at the early stages of the project and I contributed to the draft of the initial grant that financed this study. I built different models such as the Principal Component Analysis (PCA) model with initial compounds data to examine if chemical properties are sufficient to select molecules with increased bioavailability. The initial models were focused to predict the likelihood of compound visibility onHigh Performance Liquid Chromatography (HPLC) and accumulation in worm. Figures 2 and 3 of the publication are adapted from my analysis and original design.

- Phedias Diamandis, **Jan Wildenhain**, Ian D. Clarke, Adrian G. Sacher, Jeremy Graham, David S. Bellows, Erick K. M. Ling, Ryan J. Ward, Leanne G. Jamieson, Mike Tyers, and Peter B. Dirks. Chemical genetics reveals a complex functional ground state of neural stem cells. *Nature Chemical Biology*, 3(5):268–273, May 2007.

  My responsibility was the analysis of the data, described in the paper methods and supplementary methods section. In addition, I was responsible for the correct representation of molecule information and the Structure Activity Relationship (SAR) analysis. I made Figures 1B, S1, S2, S3, M1, M2, M3 and M4 and Tables S1 and S2.

- Linda Ejim, Maya A. Farha, Shannon B. Falconer, **Jan Wildenhain**, Brian K. Coombes, Mike Tyers, Eric D. Brown, and Gerard D. Wright. Combinations of antibiotics and nonantibiotic drugs enhance antimicrobial efficacy. *Nature Chemical Biology*, 7(6):348–350, June 2011.

  I built the Previously Approved Drugs (PAD) library used in this study (the components of the library are described in Chapter 3.1.1) and the supplementary files of the paper.

- Elke Ericson, Marinella Gebbia, Lawrence E Heisler, **Jan Wildenhain**, Mike Tyers, Guri Giaever, and Corey Nislow. Off-target effects of psychoactive drugs revealed by genome-wide assays in yeast. *PLoS genetics*, 4(8):e1000151, 2008.

My Bachelor thesis was focused on approved drugs to treat neurological diseases to find structural pattern that cause adverse effects. I suggested a research project to Guri Giaever that studies psychoactive drugs using the HaploInsufficiency Profiling (HIP)/HOmozygous deletion Profiling (HOP) platform to relate their adverse effects with gene deletion sensitivity. I took part in the design of the study, performed the chemical property analysis (shown in Figure 3) and processed the HIP/HOP data. This study is described in Chapter 3.

- Maureen E. Hillenmeyer, Eula Fung, **Jan Wildenhain**, Sarah E Pierce, Shawn Hoon, William Lee, Michael Proctor, Robert P. St Onge, Mike Tyers, Daphne Koller, Russ B. Altman, Ronald W. Davis, Corey Nislow, and Guri Giaever. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science (New York, N.Y.)*, 320(5874):362–365, April 2008.

  My contribution to this work was the annotation and curation of the compound data and the analysis of reproducible molecule characteristics in relation to gene deletion sensitivities.

- **Jan Wildenhain**, Nicholas Fitzgerald, and Mike Tyers. MolClass: a web portal to interrogate diverse small molecule screen datasets with different computational models. *Bioinformatics (Oxford, England)*, 28(16):2200–2201, August 2012.

  I designed the study, developed the majority of the software and wrote the manuscript (http://sysbiolab.bio.ed.ac.uk/molclass/). A detailed critical review on MolClass can be found in Chapter 4.

A.2 Additional Publications

- Philippa M. Beard, Samantha J. Griffiths, Orland Gonzalez, Ismar R. Haga, Tali Pechenick Jowers, Danielle K. Reynolds, **Jan Wildenhain**, Hille Tekotte, Manfred Auer, Mike Tyers, Peter Ghazal, Ralf Zimmer, and Jürgen Haas. A loss of function analysis of host factors influencing *Vaccinia virus* replication by RNA interference. *PLoS ONE*, 9(6):e98431, June 2014.

  A brief critical review of this work and my contributions can be found in Chapter 2.2. In brief, I imaged the screening plates on the Opera instrument, wrote the segmentation algorithm to quantify the acquired image data and carried out the statistical analysis.

- Aditi Bunker, **Jan Wildenhain**, Alina Vandenbergh, Nicholas Henschke, Joacim Rocklöv, Shakoor Hajat, and Rainer Sauerborn. Effects of Air Temperature on Climate-Sensitive Mortality and Morbidity Outcomes in

the Elderly; a Systematic Review and Meta-analysis of Epidemiological Evidence. *EBioMedicine*, February 2016.

I tutored and supported the development of the analysis workflows and drafted the R scripts for the meta-analysis. I co-wrote parts of the manuscript that explain the statistical methods and the rationale on how the publications in the meta-analysis were selected.

- Emilio Carrillo, Giora Ben-Ari, **Jan Wildenhain**, Mike Tyers, Dilon Grammentz, and Traci A. Lee. Characterizing the roles of Met31 and Met32 in coordinating Met4-activated transcription in the absence of Met30. *Molecular biology of the cell*, 23(10):1928–1942, May 2012.

  I tutored and performed all the microarray analysis and wrote the workflows to provide software that could be used for independent microarray analysis and interpretation in future.

- Alexander Ermakov, Steve Pells, Paz Freile, Veronika V Ganeva, **Jan Wildenhain**, Mark Bradley, Adam Pawson, Robert Millar, and Paul A. De Sousa. A role for intracellular calcium downstream of G-protein signaling in undifferentiated human embryonic stem cell culture. *Stem cell research*, 9(3):171–184, November 2012.

  I undertook the image segmentation and statistical data analysis for this high content small molecule screen designed to identify modulators of human embryonic stem cell differentiation. The specific goal of the screen was to characterise the role of G protein-mediated signal transduction in maintaining human Embryonic Stem Cells (hESC)s in an undifferentiated state. Drugs and ligands known to affect G protein signal transduction pathways were tested for dose-dependent effects on hESC self-renewal. I wrote scripts to quantify fluorescence of Octamer-binding transcription factor 4 (OCT4) and Nanog Homeobox transcription regulator (NANOG) proteins that are indicative of the differentiation status of hESCs and analysed screen data to classify hit compounds. Further, I wrote an Acapella script and performed the statistical analysis to quantify levels of 5-Methylcytosine (5-mC) DeoxyriboNucleic Acid (DNA) and 5-Hydroxymethylcytosine (5-hmC) DNA using targeted antibodies with an Alexa Fluor 555 (AF-555). This analysis identified a role for G-proteins and calcium signalling in human pluripotent stem cell self-renewal. These findings can be explored to maintain undifferentiated hESCs in long-term culture.

- Hironori Ishizaki*, Michaela Spitzer*, **Jan Wildenhain**, Corina Anastasaki, Zhiqiang Zeng, Sonam Dolma, Michael Shaw, Erik Madsen,

Jonathan Gitlin, Richard Marais, Mike Tyers, and E. Elizabeth Patton. Combined zebrafish-yeast chemical-genetic screens reveal gene-copper-nutrition interactions that modulate melanocyte pigmentation. *Disease models & mechanisms*, 3(9-10):639–651, September 2010.

I provided tools and analysed subsets of the data presented in this work. I contributed the chelation models for the compound classes that lead to the white fish and no-pigmentation phenotype (see Table 1) and provided Figure 4B. Further, I contributed to the HOP gene deletion analysis.

- Michaela Spitzer*, Emma Griffiths*, Kim M. Blakely, **Jan Wildenhain**, Linda Ejim, Laura Rossi, Gianfranco De Pascale, Jasna Curak, Eric Brown, Mike Tyers, and Gerard D. Wright. Cross-species discovery of syncretic drug combinations that potentiate the antifungal fluconazole. *Molecular systems biology*, 7(1):499–499, 2011.

  I contributed the statistical methods for normalisation and hit-calling as shown in Figure 1A, drafted the algorithm and statistics for the data shown in Figure 5 to rationalise synergistic interactions through integration of chemical-genetic and genetic interaction networks. I developed the application used to produce Figures 6C and 6D to test for synergistic activity (http://shiny.chemgrid.org/checkerboardr/).

- Michaela Spitzer, **Jan Wildenhain**, Juri Rappsilber, and Mike Tyers. BoxPlotR: a web tool for generation of box plots. *Nature Methods*, 11(2):121–122, January 2014.

  I initiated the contact with Nature Methods, contributed to the figure, the manuscript and setup bash scripts and optimized the server to provide a fail-proof BoxplotR service.

- Grant S. Stewart, Stephanie Panier, Kelly Townsend, Abdallah K. Al-Hakim, Nadine K. Kolas, Edward S. Miller, Shinichiro Nakada, Jarkko Ylanko, Signe Olivarius, Megan Mendez, Ceri Oldreive, **Jan Wildenhain**, Andrea Tagliaferro, Laurence Pelletier, Nadine Taubenheim, Anne Durandy, Philip J. Byrd, Tatjana Stankovic, A Malcolm R. Taylor, and Daniel Durocher. The RIDDLE syndrome protein mediates a ubiquitin-dependent signaling cascade at sites of DNA damage. *Cell*, 136(3):420–434, February 2009.

  I contributed to the initial screen that discovered E3 ubiquitin-protein ligase, Ring Finger Protein 168 (RNF168) that helped to decipher the genetics of the Radiosensitivity, Immunodeficiency, Dysmorphic, Difficulties LEarning (RIDDLE) syndrome, for details please refer to Chapter 2. I prepared the data for Figure 1B, S1B and Table S1.

- **Jan Wildenhain** and Edmund J. Crampin. Reconstructing gene regulatory networks: from random to scale-free connectivity. *IEE Proceedings - Systems Biology*, 153(4):247–256, July 2006.

  I contributed to the design of the study, wrote the network generation algorithms, built a differential equation and perturbation analysis suite in MATLAB and adapted the algorithms to analyse gene expression and time series data.

- **Jan Wildenhain**, Michaela Spitzer, Sonam Dolma, Nick Jarvik, Rachel White, Marcia Roy, Emma Griffiths, David S. Bellows, Gerry D.Wright, and Mike Tyers. Prediction of Synergism from Chemical-Genetic Interactions by Machine Learning. *Cell Systems*, 6:383–395, December 2015.

  Contributed to the design of the study, performed the analysis, wrote the software, generated the figures and wrote the majority of the manuscript.

- Andrew G. Winter, **Jan Wildenhain**, and Mike Tyers. BioGRID REST Service, BiogridPlugin2 and BioGRID WebGraph: new tools for access to interaction data at BioGRID. *Bioinformatics (Oxford, England)*, 27(7):1043–1044, March 2011.

  I developed the initial software to convert BioGRID interaction data into Proteomics Standards Initiative - Molecular Interaction (PSI-MI) eXtensible Markup Language (XML) format. I contributed to the database design for the BioGRID REpresentational State Transfer (REST) service, optimised the database and indexes and wrote the test environment BioGRID WebGraph (http://chemgrid.org/tools/).

A.3 Submitted Publications

- **Jan Wildenhain**, Michaela Spitzer, Sonam Dolma, Nick Jarvik, Rachel White, Marcia Roy, Emma Griffiths, David S. Bellows, Gerry D.Wright, and Mike Tyers. Systematic chemical-genetic and chemical-chemical interaction datasets for prediction of compound synergism. *Submitted to Scientific Data*